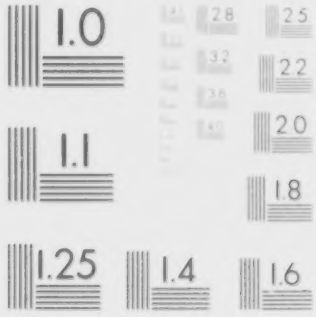
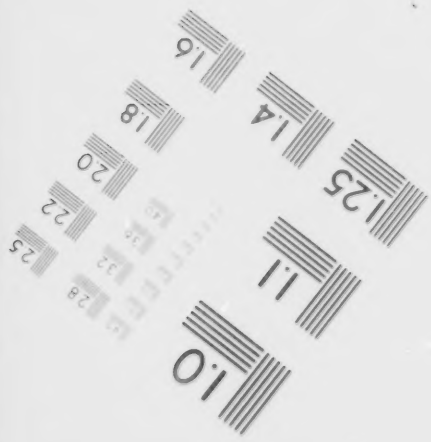
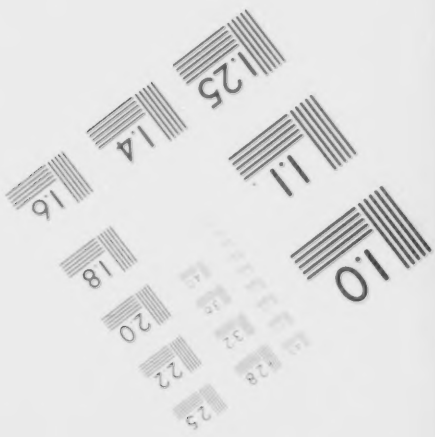
AIIM

Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910
301/587-8202

Centimeter

Inches





MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

DOCUMENT RESUME

ED 461 667 TM 033 457

TITLE	Maryland School Performance Assessment Program (MSPAP), 1998. Technical Report.
INSTITUTION	Measurement Inc., Durham, NC.; Maryland State Dept. of Education, Baltimore.; CTB / McGraw-Hill, Monterey, CA.
PUB DATE	1999-05-20
NOTE	99p.
AVAILABLE FROM	For full text: http://marces.org/mdarch/home.htm.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Program Implementation; Reliability; Scoring; *State Programs; Tables (Data); *Test Construction; Test Content; *Testing Programs; Validity
IDENTIFIERS	*Maryland School Performance Assessment Program

ABSTRACT

Maryland School Performance Assessment Program (MSPAP) assessments are criterion-referenced performance tests designed, developed, and implemented by the Maryland State Department of Education in collaboration with classroom teachers and other Maryland educators. MSPAP is the major strategy for implementing Maryland's educational reform initiative. It provides information relevant to assessing school performance and guiding school improvement plans and activities. The primary focus of the information from the MSPAP is schools, although information about individual students is available. In June and July 1998, approximately 188,000 student answer books were scored. This technical report contains information about: (1) test development; (2) test administration; (3) scoring; (4) special issues related to mathematics, algorithmic scoring, and student participation in MSPAP; (5) scaling and equating; (6) reliability; (7) validity; (8) score interpretation; and (9) the MSPAP score reports. Three appendixes contain test maps, information on the number of items comprising each outcome, and scale score ratings from each MSPAP proficiency level. (Contains 29 tables and 29 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 461 667

TECHNICAL REPORT

1998 Maryland School Performance Assessment Program (MSPAP)

Maryland State Department of Education
CTB McGraw-Hill
Measurement Incorporated

May 20 1999

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
M. Moody
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM033457

BEST COPY AVAILABLE

INTRODUCTION 6

TEST DEVELOPMENT 7

TEST ADMINISTRATIONS 11

SCORING 13

Quality Control 15

Conclusion 17

SPECIAL ISSUES 17

Mathematics 17

Algorithmic Scoring 18

Student Participation in MSPAP 18

SCALING AND EQUATING 19

Equating the Content Area Scores Across Clusters 23

Rater Year Effects Study 24

Equating 1997 and 1998 Scale Scores 26

Comparison of 1997 and 1998 Mean Scores 29

RELIABILITY 30

Coefficient Alphas 30

Standard Errors of Measurement for Proficiency Level Cut Scores 30

VALIDITY 31

Between Content Area Correlations 31

Between Content Area Correlations at the School Level 31

Test Difficulty Concerns 31

Content Validity Evidence 32

Outcomes Coverage....32
Face Validity Evidence....33
Construct Validity....33
Statistical Test Bias....33
Consequential Validity Evidence....35
Conclusion....35

SCORE INTERPRETATION....35

Scale Scores....36
Proficiency Level Descriptions....36
School Performance Standards....37
Individual Student Scale Scores....39
Outcome Scores....39

MSPAP SCORE REPORTS....41

REFERENCES....42

TABLES....45

TABLE 1....47
TABLE 3....49
TABLE 4....50
TABLE 5....51
TABLE 6....52
TABLE 7....53
TABLE 8. Summary Findings from Calibrations....55
TABLE 9. Detailed Findings from Calibration for Clusters with Choice Sets in Reading and Writing58
TABLE 10. Cluster Equating Results....59
TABLE 11....62

Rater Year Effects Study Performance ($97SS_{97}$) of State Sample on 1997 MSPAP....62

TABLE 12....63

Rater Year Effects Study Raw Score Comparisons....63

TABLE 13....64

1992, 1993, 1994, 1995, 1996, 1997, and 1998 Rater Year Effects Studies:....64

TABLE 14....65

Rater Year Effects Study Transformation Values....65

TABLE 15....66

Performance of State on 1997 MSPAP and 1998 Eguating Sample on 1997 MSPAP....66

TABLE 16....67

Equating Study Transformation Values....67

TABLE 17....68

Comparison of 1997 and 1998 MSPAP Performance by Grade and Scale....68

TABLE 18. Coefficient Alpha for 1998 MSPAP Content Areas....69

TABLE 19. Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 3....70

TABLE 20. Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 5....71

TABLE 21. Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 8....72

TABLE 22. Between Content Area Scale Score Correlations for Grade 3....73

TABLE 23. Between Content Area Scale Score Correlations for Grade 5....74

TABLE 24. Between Content Area Scale Score Correlations for Grade 8....75

TABLE 25. Between Content Area Scale Score Correlations at School Level for Grade 3....76

TABLE 26. Between Content Area Scale Score Correlations At School Level for Grade 5....77

TABLE 27. Between Content Area Scale Score Correlations at School Level for Grade 8....78

TABLE 29. Outcome Difficulty Indicators for each Grade for the 1998 MSPAP....80

APPENDIX A 81

TEST MAPS FOR 1998 MSPAP 81

APPENDIX B 82

NUMBER OF ITEMS COMPRISING EACH OUTCOME FOR 1998 MSPAP 82

APPENDIX C 83

SCALED SCORE RANGES FOR EACH PROFICIENCY LEVEL IN MSPAP` ...83

TECHNICAL REPORT
1998 Maryland School Performance Assessment Program (MSPAP)

Maryland State Department of Education
CTB McGraw-Hill
Measurement Incorporated

May 20 1999

Introduction

Maryland School Performance Assessment Program (MSPAP) assessments are criterion referenced performance tests designed, developed, and implemented by the Maryland State Department of Education (MSDE) in collaboration with classroom teachers and other Maryland educators. MSPAP is the major strategy for implementing Maryland's reform initiative and provides information relevant to assessing school performance and guiding school improvement plans and activities. The primary focus of the information provided from MSPAP assessments is *schools,* although information about individual student performance is also available.

Each May since 1991, MSPAP has been administered to Maryland students in grades 3, 5, and 8. Each student participates in nine hours of testing (reading, writing, language usage, mathematics, science, and social studies) over a five-day period, approximately one hour and 45 minutes of testing time per day. The assessments are based on the Maryland Learning Outcomes (available from the Maryland State Department of Education) that were adopted by the Maryland State Board of Education in 1990.

MSPAP is comprised of three test forms, or clusters, and one equating form or cluster from the previous year's test per grade (e.g., 3A, 3B, 3C, and 3E). Clusters are non-parallel test forms because content areas are spiraled throughout each cluster. For example, in social studies, *Peoples of the Nation and the World, Geography*, and *Economics* might be assessed in one cluster; *Political Systems, Peoples of the Nations and the World*, and *Economics* in another cluster; and *Political Systems, Geography, and Peoples of the Nations and the World* in the third cluster. Each test form or cluster assesses a combination of reading, writing, language usage, science, social studies, mathematics content and mathematics process.

Students are randomly assigned to testing groups. Random testing groups help to ensure that groups of students assigned to take each test cluster are heterogeneous in ability. In addition, random testing groups minimize influences on student performance that may occur when students are assessed in intact classroom groups by their regular classroom teachers.

Test clusters are assigned randomly to testing groups within schools and across schools in each school system and the state. Local Accountability Coordinators (LACs) implement a simple procedure (spiraling) to ensure this random assignment. Spiraling also ensures that the numbers of clusters administered within each school system and across the state will be nearly equivalent, and that schools with only three testing groups will always be assigned each of the three clusters. The Maryland State Department of Education's (MSDE's) Assessment Office approves final cluster assignments.

MSPAP is equated across years through random equivalent groups and equating clusters. Equating clusters are assigned to a representative sample of schools that have four or more testing groups in a grade and that were not used in the previous year's equating sample. Each equating cluster is given a test from the previous year's MSPAP administration so that the current year's test can be adjusted for difficulty.

Test Development

MSPAP assesses school performance on the Maryland Learning Outcomes through assessment tasks--collections of inter-related assessment activities or "items" that are organized around a theme (e.g., *Recycling* or *Salinity*). Tasks require students to respond to questions or directions that lead to a solution of a problem, a recommendation or decision, or an explanation or rationale for the responses. Some tasks assess one content area; other tasks assess multiple content areas. Activities comprising the tasks may be group or individual activities; hands-on, observation, or reading activities; and/or activities that require extended written responses, limited written responses, lists, charts, graphs, diagrams, webs, and/or drawings.

Test development consists of five phases: planning, design, development, review and revision, and field testing followed by further revisions.

Planning. MSDE instructional and assessment staff select tasks from previous MSPAP administrations to be reused. Staff then determine the learning outcomes needed to complete test clusters and plan new tasks to assess the outcomes. Up to 50% of the test may consist of reused or rolled over tasks.

Design. MSDE instructional staff write task outlines comprised of a topic area, the time allotted for the task, and the outcomes to be assessed. They design calendars showing the types of test activities and the balance of content areas for each day of testing.

Development. Approximately 170 Maryland teachers across grades 3, 5, and 8 are recruited, screened, and hired by MSDE to write MSPAP tasks and activities; develop scoring tools; and write test administration directions. Task writers are given specifications for the content areas and outcomes to be assessed; the numbers of

assessment activities per outcome and task; and the background reading materials to be used in the assessment.

Task writers are trained on the principles of performance assessment, characteristics of MSPAP, bias and sensitivity issues, and Maryland Learning Outcomes. They receive information on scoring, measurement, and administration issues; and guidelines for developing graphics and selecting tools and materials. Task writers also receive concentrated training in the areas for which they are responsible: task writing, scoring, or test administration.

Task writers develop drafts of tasks to which reading and writing cues and prompts are added where appropriate. MSDE specialists and task writers participate in an extended review and revision process that includes raising questions and resolving issues and concerns about the tasks.

One characteristic of MSPAP is the use of authentic texts. Local school media specialists select reading materials in topic areas, and reading content area staff review the materials for bias, sensitivity, and readability. After third and fifth grade "average readers" read the materials with the state reading specialist, an analysis is conducted to determine if the readability is appropriate. Only materials that average readers can read independently and show evidence of construction of meaning are used in MSPAP.

Task writers select materials, from the samples provided by media specialists, that can be used in their entirety. Occasionally, the publisher/copyright owner will not grant permission to use a text or material, and the task must be altered to accommodate other materials. For the 1998 MSPAP, MSDE secured copyright permission for 91 texts and materials.

After tasks have been drafted, they are examined to see that all activities provide a measure of the intended outcomes. Draft scoring tools, answer cue information, and sample responses are then developed. MSDE specialists and staff from the scoring contractor for MSPAP (Measurement Incorporated) review draft scoring tools and test booklets (*Answer Books, Resource Books*, and *Examiner's Manuals*) to identify problems. They then make revisions where necessary.

Review and Revision. MSPAP tasks are reviewed for:

- technical soundness,
- feasibility,
- controversial and sensitive topics,
- developmental appropriateness,
- scorability, and
- clarity.

Assessment specialists conduct technical reviews that include verifying the numbers of outcome measures in a content area and test cluster and the independent responses in a content area. At least eight independent outcome measures for each content area in each cluster are needed for scaling purposes. Four measures for each outcome measured in a cluster are needed to calculate outcome scores. The test design specifies that an outcome be measured in at least two clusters within a grade.

Local Accountability Coordinators (LACs) and assessment staff conduct feasibility reviews that include examining tasks for:

> Timing - Is adequate time allotted to tasks? Are the time blocks listed correctly in test materials?
>
> Ease of Administration - Can tasks be administered by all teachers using the same directions?
>
> Setting - Will all classrooms accommodate the administration of each task?
>
> Clarity and Complexity of Directions - Are directions clear and concise?
>
> Cluster Balance - Are content area tasks evenly distributed throughout the week? Are task varied within a day?
>
> Formatting - Is there adequate student response space in the *Answer Book*?
>
> Tools and Materials - Are materials appropriate? Adequately described? Feasible to administer? Cost effective?

Assessment and content staff conduct controversial and sensitive topic reviews in which they examine tasks for controversial language, stereotyping, and treatment of minorities, genders, and persons with disabilities. To ensure that MSPAP is free from controversial and sensitivity topics, task writers use *Guidelines to Avoid Bias and Sensitivity* that were adapted from *Bias Issues in Test Development* published by the National Evaluation System, Inc. (National Evaluation System, 1991). During the 1998 editorial review, the editors of CTB McGraw-Hill reviewed MSPAP for biased and sensitivity following the Macmillan/McGraw-Hill publication guidelines (Macmillan/McGraw-Hill, 1993).

Third and fifth grade teachers, educational psychologists, and early learning university faculty conduct developmental appropriateness reviews, to ascertain that assessment tasks are developmentally appropriate for the grade level in which they are to be administered.

Assessment specialists and experienced MSPAP scoring leads conduct scorability reviews

to verify that tasks are scorable and that they yield meaningful measures of what students understand and are able to do. Outcome/activity matches, that identify the outcome(s) being assessed by each activity, are verified.

Content specialists conduct clarity reviews to confirm that tasks are clearly written.

After MSPAP tasks have been reviewed, they are organized into an *Answer Book,* a *Resource Book,* and an *Examiner's Manual* for each grade and cluster (3A, 3B, 3C; 5A, 5B, 5C; 8A, 8B, 8C). All test booklets are then reviewed and edited for consistency, accuracy, organization, and comprehension.

Role playing is conducted to ensure that directions and timing are clear and correct. One MSDE specialist is the "teacher" and the other is the "student" who use the *Answer Book, Resource Book,* and *Examiner's Manual* as if they were taking the test. This mock administration allows for cross checking of all materials the students and test administrator will need during the actual test administration.

Field Testing. A field test is conducted to collect information on the feasibility of conducting tasks in a classroom setting, clarity of directions to students and examiners, reliability of tools and materials, and timing and scorability of tasks.

In October 1997, schools in the Inter-borough School District in southeastern Pennsylvania administered the 1998 MSPAP field test. The schools were chosen because their student populations closely matched Maryland's population with respect to race/ethnicity and gender. In addition, reading/writing instruction, collaborative learning, and hands-on learning were part of daily instruction. All new tasks appearing on the 1998 assessment were administered to two classrooms, each containing 25 to 30 students.

Observers from Maryland monitored the testing process to determine whether timing, directions, questions, or materials needed to be revised. As a result of field test administrative and scoring feedback, some tasks were slightly revised to correct timing, directions, and confusing questions. After the revisions were made, a post field test meeting confirmed that the test was ready for the May 1998 administration. Additional information may be obtained from MSDE (Westat, 1998).

Field test responses also helped to identify possible anchors (range finding), training, and qualifying responses for use in scoring training. These sample responses were selected to represent all score points possible and were based on exact agreement after discussion. (Additional sample responses for scorer training were selected from live responses "hijacked" after the MSPAP operational administration in May 1998.)

Development of Scoring Training Materials. Following field test scoring, the scoring contractor reviewed and revised scoring tools, answer cues, and sample responses to

create scoring guides for each task. Each activity was presented, followed by the scoring tool and answer cue information (typical response content, key ideas, etc.). Sample responses were selected to illustrate each score point. In the few instances in which field test scoring had not yielded any samples at a given score point, a teacher-developed sample response was utilized. Responses from the May 1998 administration supplemented these teacher-developed samples. Scoring guides were task-specific, with the exception of language in use. This generic guide was used for anchor responses to a wide array of language usage items.

The scoring contractor's senior staff developed detailed annotations to assist the Maryland-based scoring team coordinators and team leaders to train their teacher teams on scoring MSPAP. In addition, supplementary guides dealing specifically with poetry were developed to assist the expressive writing teams to apply the genre-general rubric to this particular expressive form.

Preparation of Scoring Training Materials. Training materials (training and qualifying sets) were prepared using field test and operational responses. Training sets were used for instruction and practice in task scoring. Qualifying sets were used to test the readers' ability to score accurately and to supplement the training provided by the training sets. These sets included responses from all activities to be scored by the team and were formatted to resemble the portion of the *Answer Book* that the team would score. Work was also begun on the accuracy sets that would be used twice a week during scoring to diagnose and prevent individual and/or room-wide drift away from scoring criteria. These sets closely resembled the qualifying sets described above. Preparation of training materials continued to mid June, when training began.

Pre-Packaging of Manipulatives. Tools and manipulatives for hands-on activities are pre-packaged for each testing group and its examiner through contractors. The materials are delivered to elementary and middle schools in school systems electing to use the service. When possible, materials are pre-cut or pre-measured, such as the amount of detergent or soil, and packaged for each student or teacher.

Test Administrations

Each May, the tests are administered in Maryland elementary and middle schools—to fifth grade students each morning of the first week; to third and eighth grade students each morning of the second week.

When tests are delivered to schools, they are signed for, inventoried, and immediately placed in secure storage. Two weeks prior to testing, school test administrators review test materials (*Examiners Guide, Answer Book, and Resource Book*) for only the cluster they will administer.

MSPAP is a performance assessment that requires students to produce individual responses to questions designed to elicit a variety of answers based on various kinds of information and presented in diverse ways. Responses might involve writing a few words, writing sentences, making lists, writing essays, sketching drawings, or creating tables or graphs.

Students use two booklets in taking the test: a *Resource Book* and an *Answer Book*.

The *Resource Book* contains supplementary or resource materials, such as stories, maps, charts, or other information a student needs to complete test activities. There are three versions of the *Resource Book* for each grade level, one for each form of the test.

Students also use an *Answer Book* that contains test questions and space for recording responses. For some items, students use information in the *Resource Book* to work in small groups on "pre-assessment" or group activities to help them focus on a test question. Group interaction ends before students begin work in their *Answer Books*, which is always done individually. Pre-assessment activities set the context for a test item, but do not cue or provide an answer.

Teachers use an *Examiner's Manual* to administer each form of the test. The *Examiner's Manual* contains specific instructions on how to administer each MSPAP task during the entire five-day testing period. The *Examiner's Manual* is a script that clearly tells the test examiner exactly what to say and do to move students through the test. It does not allow a test examiner to improvise in providing directions nor to provide examples unless such examples are included in the script. The purpose is to allow all students a fair chance by standardizing the way the test is given in all schools throughout the state.

Test Administration and Coordination Manual. *A Test Administration and Coordination Manual* provides information on test security and on specific test procedures to Local Accountability Coordinators who are responsible for test administration in local school systems. MSDE trains Local Accountability Coordinators in test administration. They, in turn, provide training to school test coordinators who are responsible for test administration in schools. School test coordinators train the teachers who will administer the test.

Eligible school test examiners are state-certified academic, special education, gifted and talented, English as a Second Language (ESL), and Chapter 1 classroom teachers. Test examiners are responsible for the smooth and standardized test administration and the protection of secure test materials. School staff not eligible to serve as test examiners may provide assistance during test administration as proctors only. Proctors assist the test examiner with the distribution and collection of testing materials and monitor the testing behaviors of students by keeping them on task. Proctors may not have access to secure test materials.

Participation of all grade 3, 5, and 8 students in MSPAP, except those excused or exempted according to MSDE policy, is mandatory. MSDE's policy of mandatory participation is supported by compulsory school attendance law and State Board of Education regulations on public school standards.

MSPAP Observations. In May 1998, MSDE staff observed the MSPAP administration to see how teachers, school staff, and students responded to tasks and to gather information on the administration. Test examiners submitted comments about the test on a "Concerns or Comments on the Administration of the MSPAP" form. Some examiners made general comments; others commented on specific tasks. After taking the test, each student completed a "Student Survey Form" that elicits information on whether and how the classroom instruction he/she has received is related to the areas tested on the MSPAP. Since some tasks will be reused in the next year's administration, comments were reviewed in MSDE roundtable discussions. Based on the comments and concerns of test administration observations and feedback from teachers and students, tasks are adjusted as necessary before they are administered again.

After the test has been administered, all test booklets and materials are returned to the test contractor in the same boxes in which they arrived. All scrap materials are destroyed.

Scoring

Four teams of Maryland teachers scored the assessment activities in each test form at each of the three grades using scoring guides developed by Measurement Incorporated (MI) project staff, scoring tools generated by Maryland educators, and selected sample responses chosen by Maryland educators. Each team scored the open-ended student responses and assigned the appropriate score point on a customized scan sheet. During June and July 1998, *Student Answer Books* for approximately 188,000 students were scored.

The four school sites and scoring assignments for 1998 were:

Clusters 3A and 8A: Mattawoman Middle School, Charles County Public Schools, Waldorf

Cluster 5A: Grasonville Elementary School. Queen Anne's County Public Schools, Grasonville

Clusters 3B, 5B, and 8B: Western School of Technology and Environmental Sciences, Baltimore County Public Schools, Baltimore

Cluster 3C, 5C, and 8C: Chesapeake High School, Baltimore County Public Schools, Baltimore

All booklets for a given grade/cluster were scored at the same site due to measurement implications of a multi-site model, as investigated by MSDE staff.

From previous assessments and developmental administrations of various 1998 assessment items (e.g., field test), MSDE and MI staff estimated that it would take approximately 25 minutes of reader time to score all scorable units in the answer booklet for each of the 3 clusters at each of the 3 grades—for each of the 9 grade/cluster combinations.

So that scoring loads were reasonable, the scorable units within each of the 9 grade/cluster combinations were distributed across 4 teams. At the eighth grade, a team for each of the four content areas (mathematics, science, social studies, and reading/writing/language usage) scored within their subject areas to the greatest degree possible. Each team scored assessment activities within one primary content area, although content area integration sometimes required that teams address multiple content areas. When integration occurred, enhanced training ensured accurate score decisions by all team members. Additionally, teams were selected to provide a good "fit" with the content areas being addressed by the task(s) being scored by a team. For example, a reading/science task would be predominately scored by a team of science and English/language arts specialists.

At grades 3 and 5, where most teachers work across subject areas, it was not considered crucial that each scoring team score items in only one content area. It was important to attempt to equalize reader scoring time per team, and to ensure that no one team was responsible for too many items requiring mentally demanding, complex thought processes, which might negatively affect the accuracy of readers and teams due to mental fatigue.

Staffing and Reader Distribution Throughout Scoring Sites. For each grade and cluster, four teams scored a unique set of MSPAP items--a total of 12 teams per grade and 36 teams across three grades. For each team, the data processing contractor provided a customized answer sheet. Each student's answer booklet had four customized answer sheets included with it when delivered to the scoring site.

Based upon six years of experience, MI project management established a target of 744 readers to score the 1998 MSPAP assessment, with each reader working 18 to 20 days after 2 to 3 days of training and qualifying. The number of readers required for each team varied depending upon the estimate of the relative scoring time per customized answer sheet after the 36 teams had been created. The average number of readers per team was 21. However, team size varied from 13 to 29 readers distributed across sites, grades and clusters as shown in Table 1.

See Table 1

Two leadership positions were assigned to each scoring team: a Scoring Coordinator and a Team Leaders Scoring Coordinators received five days of training by MI Project Leaders to prepare them for training readers (scorers) on their teams, monitoring readers for quality and production during the scoring process, and administering scoring in concert with MI project staff. Team Leaders, who assisted Scoring Coordinators, received three days of training.

Quality Control

Scoring accuracy is maintained by: check sets, accuracy sets, spot checks, and retraining.

Check sets, covering all MSPAP tasks, were administered on Monday mornings to help Scoring Coordinators and Team Leaders determine whether individual readers and the team of readers were continuing to score accurately and consistently, especially on items that were complex and difficult to score. As scoring progresses, readers may "drift" away from score points, especially after a weekend away from scoring. As inconsistencies and inaccuracies were detected, Scoring Coordinators and Team Leaders held discussions with the team and assisted individual readers to improve accuracy.

Accuracy sets. Accuracy sets were administered on Tuesday and Thursday mornings to determine whether teams of readers maintained appropriate levels of accuracy during the scoring process. Therefore, each accuracy set included a student response for each scorable unit, and each reader's average score was recorded so that the mean score for each accuracy set could be calculated. These mean scores were used to construct Tables 2 through 7, which will be used to analyze quality control for this scoring project.

Readers in 35 of the 36 teams were given at least 5 accuracy sets, usually 6 to 7 sets. Readers who scored below 70 percent on any accuracy set received additional training immediately from the Scoring Coordinator or the Team Leader and were released from retraining only after the leaders determined that scoring problems were resolved. If the scoring problems were not resolved, the reader was dismissed from the scoring project.

In *spot checking*, a Scoring Coordinator or Team Leader rescored a booklet to estimate a reader's overall accuracy, to determine specific items with which a reader was having difficulty, or to ascertain specific items that were causing individual readers to perform poorly on check sets or accuracy sets.

In *retraining*, Team Leaders or highly accurate readers used the scoring guide and student papers to assist readers who had experienced problems maintaining appropriate accuracy

levels. Small groups of readers who shared a common scoring difficulty were also retrained to improve their scoring accuracy.

Reader accuracy results. In 1998, 206 accuracy sets were administered across all 36 scoring teams. The reader accuracy set mean scores for each scoring team are shown in Tables 2, 3, and 4 for grades 3, 5, and 8 respectively.

See Tables 2-5

The results are summarized in Table 5 by grade and across all three grades. The results are reasonable and acceptable for scoring open–ended performance assessment items. Forty-five percent of the sets had mean scores between 80 to 89% and 36 percent were at or above 90% accuracy. Thirty-five percent had mean set scores between 70 to 79%, and only three of the accuracy set mean scores were below 70% accuracy. The results for the 1998 MSPAP were similar to those for the previous three years. The accuracy set mean scores were similar to past years.

The averages across the accuracy sets for each team could be calculated because the sets contained the same number of scorable units. However, it was not possible to calculate the averages across different teams because the numbers of scorable units varied considerably from team to team. When the accuracy set mean scores were studied in terms of content area, the results were reasonably predictable yielding no major surprises.

Bearing in mind that few teams addressed only one content area, it is possible to look at results for predominant content area focus in the eighth grade. Results by content area for the eighth grade are displayed in Tables 6 and 7. From past scoring of performance assessments it was reasonably predictable that the scoring of mathematics would yield relatively higher and somewhat more consistent accuracy set scores. As commonly found in the hand-scoring of performance activities, the accuracy set mean scores for reading/writing/language usage were lower than those for mathematics, science, and social studies.

See Tables 6-7

In grades 3 and 5, the items to be scored within each content area were distributed across teams to such a degree that it was not possible to analyze accuracy set mean scores systematically by content area. Past experience in scoring open–ended performance assessment items indicated that the relationships between content area and accuracy set scores at grades 3 and 5 would be similar to those at grade 8. In addition, MI Project Leaders and the Scoring Coordinators and Team Leaders felt that it was more difficult to

train readers to score items consistently in reading/writing/language usage than in other content areas. These responses more often measure higher level skills and objectives; and they more often require holistic scoring decisions rather than more discrete decisions.

Conclusion

The factors that interacted to produce improvements in training and scoring productivity are:

Early field testing to provide an adequate time frame for scoring booklets, selecting training materials, and preparing annotated scoring guides.

An adequate time frame for planning and implementing activities for both CTB (the data processing contractor) and MI.

Increased experience of MI and Maryland project staff. Many readers and leadership staff in Maryland had not only gained another year's experience in scoring MSPAP activities, but had also become increasingly involved in other MSPAP activities, such as task development or rangefinding (field-test scoring).

Special Issues

Mathematics

Prior to the 1996 MSPAP, 13 mathematics outcomes were measured, more than twice as many outcomes as were measured in other content areas. The number of measures needed in a cluster made designing the mathematics component difficult and often made individual tasks too long. Therefore, some mathematics outcomes were combined, thereby reducing the number of mathematics outcomes to nine. All mathematics outcomes are still tested, but there are fewer mathematics measures. For example, because geometry and measurement were combined, instead of needing four measures of each outcome for reporting purposes, only four total measures are needed. The mathematics supervisors in each school system accepted this change.

The 1998 MSPAP included limited problem solving. The problem-solving outcome has been difficult to include in the test because of the scope of true problem solving. Additionally, scoring time and training needed to be slightly modified. However, it was important to include problem-solving activities because of their emphasis at the national and state levels.

Algorithmic Scoring

Prior to 1995, students who were absent on one or more days of MSPAP testing could not obtain a content area scale score if they missed any day on which the content area was assessed. Algorithmic scoring is a process for deriving a score for students who were absent, but who had 60% or more of the responses in a content area and a minimum of eight independent measures.

Algorithmic scoring uses a maximum-likelihood estimation which is a general method of finding good parameter estimates in a model. Since table scoring is based on complete score records, the ability estimates of absent students are inaccurate (underestimated). Therefore, students scored algorithmically can have their ability more accurately estimated using a maximum likelihood estimator, which approximates student ability using the data available. Beginning with the 1995 MSPAP, CTB McGraw-Hill scored all students algorithmically. (Before 1995, CTB used table scoring.)

To be eligible for algorithmic scoring, a student must have attempted at least 60% of the content area and at least eight independent items. Exceptions include the content areas of writing and language usage, as well as any "short" test. A short test is a test of fewer than eight independent items, typically math process. Since the mathematics total score is a combination of mathematics content and process, mathematics does not benefit from this scoring process. Because writing is a three-item test, if a student responds to the extended writing prompt (scored 0-3) and to one of the two limited writing prompts (scored 0-2), then a student should receive a score. (From 1992 to 1994 only one extended and one limited writing process comprised the writing test. Therefore, MSPAP added another limited writing process to the writing scale in 1995. If students missed one of the limited writing process prompts, they still received a writing score.) Language usage is the content area most vulnerable to absence vulnerability because language usage measures are captured throughout the week. Therefore, language usage is scored for absent students as long as six or more of the responses in the student's language usage vector have either valid scores or score codes. Score codes are assigned when the student response is invalid which may be blank, off-task, or unscorable response.

Algorithmic scoring increased the number of students who received at least one score. In 1998, across all grades and content areas, more than 15,000 more scores were computed using algorithmic scoring. This method of scoring gave a more accurate reflection of student performance in a school or system.

Student Participation in MSPAP

It is the policy of Maryland to include all students to the fullest extent possible in all state assessment programs. Testing accommodations that meet state guidelines are provided to

help students with disabilities and English as a Second Language (ESL) students participate more fully in assessments and better demonstrate their knowledge and skills.

MSPAP permits five categories of accommodations (scheduling, setting, equipment, presentation, and response) with 31 accommodations under the five categories for students with Individualized Education Programs (IEPs) and ESL students. Most accommodations do not invalidate student scores; however, in some cases, the student will not receive a score if the validity of the work that has been accommodated has been compromised. For example, if an examiner must read sections of the test to a student, the reading construct has been comprised. The student is not reading but listening; therefore, the student will not receive a reading score for the test. The student will, however, receive scores in all other content areas.

Students with disabilities may be exempted from MSPAP if they are not pursuing the Maryland Learning Outcomes but, instead, are pursuing alternative or life skill outcomes. ESL students may be exempted if they do not have the minimum language proficiency required for participation in MSPAP. ESL exemptions are limited to one test administration, i.e., a student exempted in grade 3 cannot be exempted again in grade 5.

Students may be excused from testing for a variety of reasons, such as demonstrating inordinate frustration, distress, or disruption of others and/or require accommodations that the school is unable to provide.

Students who are exempted do not take the test and are not included in the calculation of MSPAP scores for a school. Students who are excused do not take the test, but are included in the calculation of MSPAP scores. In other words, the school is not held responsible for students who are exempted from the test; it is held responsible for students who are excused from the tests.

Scaling and Equating

Scaling and equating MSPAP scale consists of two major phrases. In Phase I, item calibrations are conducted to obtain the item parameters for each cluster. Misfitting items are identified and removed from the scale. Cluster equating is conducted to adjust the differences in difficulty among test forms. In Phase II, the results of two studies were used to link students' performance on the 1998 scale to the 1997 score scale. The first, *Rater Year Effects Study*, was designed to determine differences between raters who scored the 1997 MSPAP and raters who scored the 1998 MSPAP. The second, *Year to Year Equating Study*, was designed to equate the scores of two samples of students who were administered the 1997 and 1998 MSPAP in 1998.

The results of the two studies were combined to produce values that could be used to transform students' 1998 MSPAP scale scores to the 1997 score scale. This

transformation permits comparisons to be made between the performance of students administered the MSPAP in 1997 and 1998.

Test Cluster Equating

To adjust for differences in difficulty among test forms, MSPAP is equated horizontally. Equivalent scores are established on test forms in a grade (e.g., Cluster 3A, 3B, 3C), but not across grades (e.g., grades 3 and 5). Therefore, MSPAP scores can be compared within a grade, but not between grades.

To equate horizontally, equivalent group design (administering tests to be equated to groups of examinees equivalent in terms of the skill measured by the tests) is used. In MSPAP, equivalent design is implemented by randomly assigning students to test groups by their Local Education Agency (LEAs). For cluster equating, at least three test groups of randomly assigned students in a grade in a school are administered one of three test clusters. Across Maryland, approximately 20,000 students in a given grade are assigned to each test cluster.

Rater-Year Effects Equating

Rater-year effect equating is conducted to determine and adjust for rater or scorer variance from one year to the next. Approximately 1,500 *Student Answer Books* per grade from the 1997 MSPAP administration were rescored by 1998 raters.

Year-to-Year Equating

To adjust for differences in difficulty from year to year, a test form from the previous year's edition is administered. For the 1998 annual equating, 2,500 students per grade were selected to take a 1997 cluster. In each school system, one or more schools were randomly selected; in each school, a test group of randomly assigned students was selected. In each school system, the number of schools chosen for equating was proportional to the system's representation in the state as a whole. Because a minimum of three test groups in each grade take MSPAP, only schools with more than three test groups in a grade were selected for equating

The next step in the equating study was to identify a group of students in each grade who took the 1998 MSPAP and who were equivalent to the 1998 group of students administered the 1997 MSPAP cluster. Following MSPAP administration, CTB counted the number of valid students from each LEA who took the 1997 MSPAP for the equating study and randomly sampled from the equating schools in the LEA the same number of students who took the 1998 MSPAP. This procedure ensured that the numbers of students from each LEA were identical in the two groups used for the equating.

The critical assumption that must be met to use the equivalent groups design is that the groups taking the tests to be equated are equivalent, not representative. CTB proportionally samples from all LEAs to construct equating groups to avoid the appearance that any undue influence on the equating results is exerted by one LEA or another.

Analysis procedure

The equating process involves constructing an equation that permits the translation of scores obtained on one test to corresponding scores on a second test. It was the responsibility of CTB to express the 1998 obtained MSPAP scores on the 1992 score scale so that performance in the test years are comparable.

The method used derives a linear equation that can be used to adjust the scores on one test so that they correspond to the scores given for comparable performance on the target test. In the case of cluster equating, this target test was the 1998 cluster that had the most regular cumulative score distribution. In the case of the 1997-1998 equating, this target was the 1997 clusters administered in 1998 for the equating study.

When tests are scaled using item response theory, it is necessary that linear equating be done. Traditionally, linear equating based on equivalent groups has involved merely equating means and standard deviations. However, considering only means and standard deviations can produce unsatisfactory equating for tests such as MSPAP that have few items or unusual score distributions. Therefore, for equating MSPAP a procedure was used that was more detailed and robust than equating means and standard deviations. This procedure, called the linear equipercentile procedure, determined the linear transformation that most closely aligned the greatest number of score points possible.

The linear equipercentile procedure had several steps. First, pairs of scores on the two tests that had the same percentile rank were identified. Then, the linear function that most accurately described this equipercentile result was determined. For the vast majority of tests, the score pairs fell on a straight line; therefore, the linear function ran through all the pairs.

As in previous years, the operating principle for equating was "the greatest accuracy for the greatest number." In other words, the equating line was located so that it passed through as many scores as possible. It was also located with attention on the Proficiency Level 3/4 cut score.

Item Set Calibrations

As in previous years, 1998 MSPAP items were calibrated separately by cluster. The calibrations for each cluster were based on stratified random samples drawn from the pool of students in the state who were administered the cluster. The strata consisted of the 24 Maryland LEAs. Within each grade, students were sampled such that their proportional representation in the calibration sample corresponded to their LEA's proportional representation in the state. Table 8 shows that the sample sizes for each calibration ranged from 7,499 students to 7,502 students. Separate samples were drawn for each set of items to be calibrated.

Table 8 shows that item calibrations, or item scalings, were carried out for reading, writing, language usage, mathematics content, mathematics process, science, and social studies. Mathematics content and mathematics process items were assigned to different scales because it was known that some of the mathematics process items would be dependent on the mathematics content responses.

Table 8 shows that no items were deleted due to group administration or at the request of MSDE prior to the initial scaling.

The Two-Parameter Partial Credit model (CTB McGraw-Hill, 1992, p. 4-4), as implemented by the PC based program PARDUX (Burket, 1992), was used for scaling the responses to the 1998 MSPAP items. Trait estimates, as well as standard errors of measurement for these estimates, were developed using the same procedures that were used in previous test editions. For two items assessing writing content, PARDUX could not provide parameter estimates. These items typically had difficulties that were extreme and different from the other items in the scale. For each of these items, plots of students' observed performance were used to fit tracelines "by hand." That is, the graphical display capability of PARDUX was used to examine observed item tracelines. Item parameters that produced tracelines that most accurately represented the observed data then were identified interactively.

The same two types of model fit analyses used to evaluate MSPAP items in the past were used again in 1998. The two types of analyses used an analogue to Yen's Q_1 (Yen, 1981) fit statistic and an analogue of Yen's Q_3 dependency statistic (Yen, 1984). The Q_1 statistic was used to compare observed and expected tracelines statistically. Also, graphical representations of these lines were examined. The Q_3 statistic was used to examine local dependence. Even though local dependence is still examined, it is important to remember that there have been no testlets of dependent items constructed since 1992.

Items with differences between students' observed and expected performance that exceeded criterion values were flagged for further study. These criterion values are described in detail in the Technical Report for the 1991 MSPAP. The items that exceeded

the criterion values used for the 1998 MSPAP are given in Table 8. Math process and reading had some items flagged for poor fit.

There are limitations to the usefulness of fit statistics such as Q_1. First, chi-square measures such as Q_1 are greatly influenced by the deviation of observations from very small expectations; this influence results in high chi-square values for deviations of no practical significance. Another limitation is that performance on an item is implicitly included in the model via the trait estimate. With shorter tests, such as math process and writing, there is substantial part-whole contamination in comparing item observed performance with predictions that implicitly include that item via that trait estimate. Lastly, the Q_1 statistic criterion is very conservative; it often flags items that in fact fit really well. Due to these limitations, the Q_1 statistic was used as a flag for potential misfit. The fit of each flagged item was then further evaluated using detailed fit information and both graphically within PARDUX.

If very large differences between students' observed and expected performance occurred on an item, the item was judged to have poor fit and was deleted. Table 8 shows that in 1998 no items were deleted due to poor fit.

When reading for literacy experience is measured, students in cluster 3A, 5B, and 8C were allowed to select from three or four passages the one they wanted to read. When writing for personal expression was measured, students in 3A, 5B, and 8C were allowed to choose their topic they wanted to write about and the form of writing they wanted to use. Table 9 details the calibration information for the reading and writing choice clusters. The writing choices of poem and play were not widely selected by students. The fit of each flagged item was then further evaluated using detailed fit information and both graphically within PARDUX. Table 9 shows that no items were deleted due to poor fit.

See Tables 8-9

Equating the Content Area Scores Across Clusters

The procedures used to equate content area scores are comparable to those used to equate content area scores of previous MSPAP forms. Specifically, cumulative scale score distributions for the calibration sample for each cluster and content area were obtained. In each grade, the content area scores of one cluster were designated as the target distribution. FLUX was used to carry out an equipercentile equating procedure to align distributions of content area scores from each of the two other clusters so that they matched the target distribution as closely as possible. A linear transformation that produced the closest alignment between the target and a non-target score distribution was identified and used to adjust the non-target scores to the score scale.

Table 10 specifies the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) for each content area and cluster. Note that the LOSSes and HOSSes are the same for the three clusters used to assess a given content area in a grade.

See Table 10

Table 10 also indicates the percentage of students in the calibration samples at the LOSS and the HOSS, which is a useful measure of floor and ceiling effects. The table shows that there are substantial floor effects in writing and language usage. These tests are uniformly difficult and very short, and many students in the calibration samples received scale scores at the LOSS.

Rater Year Effects Study

Method

For this study, the responses of approximately 1,500 randomly selected students who had taken the 1997 MSPAP (Clusters 3C, 5A, or 8A) were re-scored by raters who scored the 1998 MSPAP. The 1998 raters were trained, using Scoring Guides developed for the 1997 MSPAP, by Measurement Incorporated (MI), the hand-scoring contractor for the MSPAP in both 1997 and 1998.

Analyses

Analyses of the rater effects were conducted separately by scale within Grades 3, 5, and 8. To determine the magnitude of the rater effect for each scale, the 1997 item parameters were used to generate 1997 scale scores for the students in the study. The first set of scale scores ($97SS_{97}$) was based upon the ratings that the students received when they were tested in 1997. The second set of scale scores ($97SS_{98}$) was based on the ratings that these students received when they were re-scored by the 1998 raters. Both sets of scale scores were expressed on the 1997 score scale.

Linear equipercentile equating procedures, as implemented in the computer software program FLUX (Burket, 1992), were used to align the $97SS_{98}$s with the $97SS_{97}$s. The linear transformation that best expressed the adjustment to the $97SS_{98}$s was used to define the magnitude of the rater effect for each scale assessed in each of the three grades.

Results

Table 11 shows the mean 1997 scale scores ($97SS_{97}$) for the samples used in the Rater Effects Study and the mean scale scores for the State reported in the 1997 Forms Effects Study for Clusters 3C, 5A, and 8A. The table shows that for all three grades, the samples tended to have slightly higher scale scores than did the population of students who were administered this cluster. Overall the differences were typically less than one tenth of a standard deviation.

The average raw scores obtained in 1997 and the values obtained when they were re-scored in 1998 are given and compared in Table 12. Positive values, given in the last column of the table, indicate that the 1998 raters graded the students more leniently than did the 1997 raters, that is, they gave the students higher scores on the average. Negative values, in this column, indicate that the 1998 raters graded the students more severely than did the 1997 raters, that is, they gave the students lower scores on the average.

This table shows that the 1998 raters evaluated the samples similarly to the 1997 raters. In grade 3, the differences were less than one tenth of a standard deviation in all content areas except for Language Usage. The differences in average raw scores obtained from fifth grade samples indicate that the 1998 raters evaluated grade 5 tests more leniently in the content areas of Reading, Language Usage, Math Content, Math Process, and Social Studies, and more severely in the other two content areas. The average raw score differences demonstrate that the 1998 raters were slightly more lenient than their 1997 counterpart in evaluating the grade 8 tests of Writing, Language Usage, Math Content, and Math Process, but slightly more stringent in the other three content areas. The largest discrepancy in average raw scores across all three grades was in grade 5, Social Studies.

A comparison between the mean differences reported in the current study and those reported for 1992 through 1998 MSPAPs are given in Table 13 in terms of standardized mean differences. Positive differences indicate that the raters who scored in the year that the study was done were more lenient than the raters who scored in the previous test year. Negative differences mean that the raters who scored in the year that the study was done were more severe than the raters who scored in the previous test year.

Table 13 shows that in terms of raw scores the rater effects generally were quite small in 1998, ranging from zero to one-tenth of a standardized mean difference in either direction for all content areas in the three grades with the exceptions of Social Studies in Grade 5. The 1998 results indicate small differences between the 1997 and 1998 rater groups. The 1998 results also indicate that the 1997 and 1998 raters were not consistently more lenient or severe relative to previous study years.

The values of the multiplicative (R_1) and additive (R_2) components of the transformations that best aligned the $97SS_{98}$s with the $97SS_{97}$s are given in the first two columns of Table

14. When applied to the 1997 parameters, these values adjust the 1997 parameter values for the 1998 rater effects. To illustrate the magnitude of the adjustment, the transformation values were applied to a scale score of 500. The value of 500 was chosen because the average 1997 scale score was near 500. Since the values given in Table 14 are expressed in terms of the scale score metric, they will resemble but not mirror the raw score results given in Table 12, since raw scores and scale scores have a non-linear relationship.

See Tables 11-14

Equating 1997 and 1998 Scale Scores

Method

For this equating study equivalent groups of students administered the 1997 and 1998 MSPAP were required, since no anchor items were available to link the tests administered in the two years. Accordingly, in 1998 approximately 2,500 third grade, fifth grade, and eighth grade students were selected to take 1997 MSPAP test books in May, 1998, while their counterparts were administered the 1998 MSPAP. The third grade students took Cluster 3C from the 1997 MSPAP; the fifth grade students took Cluster 5A; and the eighth grade students took Cluster 8A. These are the same books as those that were used for the Rater Effects Study just described.

The test groups in each grade were selected using stratified random selection procedures. Following a priori decisions to involve in the study no more than one test group per school and to use only Maryland schools with more than three classrooms, schools within each LEA were randomly selected to provide test groups for the Equating Study. Schools were selected separately for Grades 3, 5, and 8. The number of schools selected within each LEA was proportional to the representation of the LEA in the state. Within each school selected to contribute a test group in a given grade, the test group was randomly selected. Since all eligible students in a grade were randomly assigned to test groups, this test group was representative of the students in the school in the grade of interest.

Students' responses to the 1997 test books were scored by the same 1998 raters who were trained to score the 1997 books for the Rater Year Effects Study. For each scale, the students were screened to ensure that they had ratings for all the items used to assess that scale in the cluster of interest.

Only those students meeting the screening criteria were used in the analyses for a given scale. For the 2,500 cases administered a 1997 cluster in each grade, Table 15 shows that the screening process left a minimum of 2,323 students per scale for the analyses.

To develop equivalent groups administered the 1998 test, it was decided a priori to select students who had been administered the clusters used as targets in the 1998 cluster equating. The target clusters typically had the most items, therefore the most reliable measurement. The target clusters also typically had smooth score distributions and items with good fit. The target clusters for the cluster equating in Reading were 3B, 5C, and 8B; for Writing-- 3C, 5C, and 8A; for Language Usage--3B, 5C, and 8B; for Math Content--3A, 5C, and 8A; for Math Process--3A, 5B, 8B; for Science--3A, 5B, and 8B; and for Social Studies--3B, 5B, and 8B.

The equivalent groups administered the 1998 target clusters in each grade were developed separately for each scale within the grade. To do this, the number of 1998 students selected from each LEA for the analyses was the same as the number of students from that LEA who took the 1997 test books for the Equating Study and had valid scores on the scale. Let's say, for example, that in the Equating Study we found that there were 24 students from LEA #1 who took 1997 Cluster 3C and had valid reading scores. To develop an equivalent group to use for the equating of the 1997 and 1998 Reading scales, we would randomly select 24 students from the same LEA who had valid scores on the 1998 target cluster--3B.

See Table 15

Analyses

The students in the Equating Study who took the 1997 test books were scored using the 1997 item parameters estimated for the items in these books. The use of these parameters ensured that these students' scale scores would be expressed in terms of 1997 scale scores; since these students' responses were scored by 1998 raters, it is useful to designate these scale scores as $97SS_{98}$. The students who took the 1998 test books were scored using the 1998 item parameters estimated for the items in these books, so that these students' scores were expressed in terms of 1998 scale scores. Since these students' responses were scored by the 1998 raters, their scale scores can be designated $98SS_{98}$. In the equating analyses, the lowest and highest obtainable scale scores from the 1997 MSPAP were used. This was done so that the scale scores for all students would not have scores that fell beyond the range of scale scores obtainable in 1997.

Equating procedures implemented by FLUX (Burket, 1992) were used to align the $98SS_{98}$s with the $97SS_{98}$s. The linear transformation that best aligned the $98SS_{98}$s with the $97SS_{98}$s was used to express the $98SS_{98}$s on the 1997 scale.

Results

It is important to emphasize that the equivalence of the two samples used in the equating is critical for the soundness of the equating. The only data available to measure the equivalence of these samples were the distributions of students across LEAs, which indicated that the equating groups matched exactly in terms of the number of students taken from each LEA.

In the paragraphs that follow, comparisons are made between the test performance of the equating samples administered the 1997 books and the state as a whole in 1997. These comparisons are useful for the purposes of documentation and general information.

Table 15 describes the sample of students' $97SS_{98}$s and compares these scores to state means estimated for 1997. In examining this table, it is important to keep in mind that the $97SS_{98}$ reflect performance on 1997 items evaluated by 1998 raters, adjusted for the differences between the 1997 and 1998 raters. In other words, these statistics reflect the scores that would have been obtained had 1997 raters been used.

The table shows that the scale scores are relatively similar across the grades when the State and the sample results are compared. For grade 3, the differences in means are less than one tenth of a standard deviation in all content areas except for Social Studies. For Social Studies, the performance of the 1998 sample on the 1997 MSPAP equating cluster (i.e., 3C) was better relative to the statewide 1997 MSPAP performance. For grade 5, the 1998 sample performed better than the 1997 MSPAP population on the average in Reading, Writing, and Language Usage. The differences in mean scale scores in the other four content areas are less than one tenth of a standard deviation. For grade 8, the mean scale scores of the State and the sample are very similar, the differences are less than one tenth of a standard deviation across all content areas. Inspection of the case counts by LEA in each grade revealed that the proportions of students from each LEA were quite similar to the proportion of students that the LEA represents in the state.

The values of the multiplicative (T_1) and additive (T_2) components of the transformations that best aligned the $98SS_{98}$s with the $97SS_{98}$s are given in the first two columns of Table 16. In addition, the result of applying these transformation values to a scale score of 500 is shown in the third and fourth columns of the table to provide a sense of the size and direction of the test effect. Positive values in the fourth column of the table indicate that a scale score of 500 obtained on the 1998 MSPAP was transformed to a score greater than 500 on the 1997 scale. Negative values indicate that a scale score of 500 obtained on the 1998 MSPAP was transformed to a score less than 500 on the 1997 scale.

See Tables 15-16

Comparison of 1997 and 1998 Mean Scores

Table 17 provides data permitting comparisons between the MSPAP performance of the students in 1997 and 1998 on the average. Both the 1997 and 1998 results reflect the average scale scores obtained by the student populations in three grades 3, 5, and 8.

The results in Table 17 suggest that there was a slight improvement in student performance in all content areas except for Language Usage and Math Process in grade 3, Math Process and Social Studies in grade 5, and Reading in grade 8.

Caution must be exercised when interpreting the differences observed in Table 17. This is especially true for the Writing and Math Process results since they were very short tests and had large standard errors. All the differences observed in the last column of Table 17 are too small to allow an interpretation of the trend of the performance of the Maryland students by themselves. However, consistently higher scores for the 1998 students suggest some degree of growth occurred in each grade for several content areas.

When considering these results it is important to remember that there are many different statistics that can be used to describe student performance. Average scores are a convenient statistic, but when distributions are as skewed as many are for the MSPAP, the median may be a better indicator of typical test performance. The reports produced by the state of Maryland summarize performance in terms of Proficiency Standards; these bands constitute another set of statistics by which performance can be described. The statistic used will affect the results one obtains and the conclusions one draws about growth or declines in performance over years. The average scores reported in Table 17 may not provide the same picture of student performance as that obtained when other statistics are used to describe this performance.

See Table 17

Review and Decision Points for the 1998 Equating. As an equating assurance check, review and decision points were examined for all cluster and annual equating. MSDE, the National Psychometric Council, and CTB McGraw-Hill reviewed the cluster scaling and equating, rater year effect equating, annual equating, and performance results before each subsequent step of the process was undertaken. Through this process the test characteristic curves and percentile rank correspondences were found to be very acceptable for the 1998 MSPAP equating.

Reliability

Coefficient Alphas

Coefficient alpha is a reliability measure suitable when items have a variety of score levels (Allen & Yen, 1979). The coefficient alphas based on the calibration sample are reported in Table 18 by grade and cluster. Refer to Table 8 and 9 for the sample sizes and the number of items comprising each scale. The alpha coefficients for each grade and content area are generally around 0.85 except for writing, which is generally around 0.70. Generally, the mathematics process scale has lower alphas than other scales as well. Both the writing test and mathematics process test are short tests, unlike mathematics content and social studies. For example, the writing test is comprised of three items spanning at least two different writing purposes, unlike mathematics which usually has more than 30 items per cluster. (For information pertaining to the number of items comprising a scale, refer to Table 8). The coefficient alphas for each MSPAP test within each cluster are consistent with other constructed response tests (e.g., see KIRIS Accountability Cycle Technical Manual, 1997).

The coefficient alphas obtained in the MSPAP writing assessment are typical of short tests. The MSPAP writing results are similar to the coefficient alphas obtained on the Maryland Writing Test (MWT), a performance assessment comprised of two items. The coefficient alphas for the MWT range from 0.50 to 0.55. Therefore, the reliabilities for the writing portion of the MSPAP are considered acceptable as well.

See Table 18

Standard Errors of Measurement for Proficiency Level Cut Scores

The standard error of measurement (SEM) is displayed in Tables 19 to 21. These SEMs are for individual scores in each content area. No test provides an exact point estimate. Instead, all scores have some degree of error. The SEM, produced through the Two-Parameter Partial Credit model, is influenced by the amount of information provided by each item and the number of items contributing to a content area. In this way it is similar to the coefficient alpha. As can be noted from the tables, SEMs are usually smaller in the middle of the scale distribution (i.e., Proficiency Level 3/4 cut) and larger at the ends (i.e., HOSSes and LOSSes). Because the SEM is a function of item and test information, higher standard errors of measurement are not surprising in writing, language usage, and math process which are all short tests of three to nine items.

See Tables 19 to 21

Validity

MSPAP validity evidence is collected to support and validate intended interpretations and uses of scores from the assessment. Additionally, it is important that MSPAP assesses the skills and knowledge that are documented in the Maryland Learning Outcomes document. The validity evidence described below is organized around these goals.

Between Content Area Correlations

Correlations were calculated to examine the relationships between the content area scale scores at each grade level. The relationships can be described as moderate to strong. In Tables 22 through 24, in third grade, the largest relationship is between mathematics and science, and the smallest is between writing and reading. In the fifth grade, the largest relationship is between mathematics and science, and the smallest is between writing and reading. In the eighth grade, the largest relationship is between language usage and writing, and the smallest is between language usage and mathematics. These findings are similar to the moderate to strong correlations found among MSPAP content area scale scores, CTBS/4, and teacher ratings calculated in a special study of the 1991 MSPAP test edition (see CTB McGraw Hill, 1992, Tables 9-8 through 9-10).

See Tables 22 to 24

Between Content Area Correlations at the School Level

Correlations were also calculated to examine the relationships between the content area scale scores at each school. The relationships can be described as strong. In Tables 25 through 27, in third grade the largest relationship is between science and social studies, and the smallest is between language usage and mathematics. In the fifth grade, the largest relationship is between science and social studies, and the smallest is between language usage and mathematics. In the eighth grade, the largest relationship is between science and social studies, and the smallest is between reading and mathematics.

See Tables 25 to 27

Test Difficulty Concerns

MSPAP was developed with standards for the year 2000. The test was built around what students are supposed to be learning. Two impacts of test difficulty are (1) the test

information function does not overlap well with student scores, and (2) higher standard errors at the lower and upper regions of the distribution. Since 1992, the fit between the test and student achievement has been improving.

Content Validity Evidence

Content validity evidence refers to the degree to which an assessment reflects the content it was designed to assess. The Maryland Learning Outcomes, the basis for learning, instruction, and MSPAP assessment activities, are based on national curriculum standards and learning theories. For example, the reading outcomes are similar to the NAEP reading assessment objectives and based on the reader response theory. Similarly, the writing outcomes are based on long-recognized modes of discourse, and the mathematics outcomes are based on the National Council of Teachers for Mathematics (NCTM) standards for curriculum and evaluation. The science outcomes are based on Project 2061 by the American Association for the Advancement of Science (AAAS). Additionally, the social studies outcomes are underpinned by the work of many groups including the Association of American Geographers, the Commission on History in the Schools, and the Joint Council on Economic Education. Moreover, the assessment tasks are developed by content area and grade specialists, specifically teachers. Each task development team is given specifications on which outcomes to assess in their task. After tasks are completed, they are reviewed.

A high degree of match between assessment activities and the outcomes they assess is ensured through multiple reviews during the development of tasks, scoring tools, and scoring guides. A task is reviewed by the task writers, test scoring teams, test administration teams, and is field tested. These reviews allow for the opportunity to confirm that the specified outcomes as defined by the Maryland Learning Outcomes document are being assessed.

Outcomes Coverage

Coverage of outcomes by assessment activities is proportionally balanced according to the relative importance of the outcomes at different grade levels. A high degree of match between assessment activities and the outcomes they assess is ensured through multiple reviews during task development and development of scoring tools and guides. All of these reviews allow for the opportunity to confirm that the specified outcomes are indeed being measured-as defined by the Learning Outcomes document. Appendix B presents the Maryland Learning Outcomes and the number of items measuring each outcome by grade and cluster for 1998 MSPAP.

See Appendix B

Face Validity Evidence

Face validity evidence refers to the accuracy with which the test appears to measure what it is supposed to measure. MSPAP has substantive face validity evidence. It is a performance-based assessment that uses authentic and real-life situations as assessment tasks. In addition, reading selections are full-length published works rather than excerpts contrived for use in a test. Furthermore, the test is administered to random groups of students who work in small groups that reflect authentic situations. MSDE content chairs assign tasks to be written for a group of outcomes.

Construct Validity

Construct validity is considered to be the unifying concept for all views and types of evidence of test score validity (see, for example, Messick, 1989, p. 13). One way to assess the construct validity of MSPAP is to compare its results with similar tests. Since MSPAP reflects the NCTM standards and the reader-response model of reading, MSPAP results can be compared to Maryland's NAEP results.

Maryland's fourth grade National Assessment of Educational Progress (NAEP) reading performance showed 26% performing at/above the "proficient" level on the 1994 NAEP Trial State Assessment. These results were similar to 1994 MSPAP reading results when 30.6% of the state's third graders and 30.2% of the fifth graders scored at the satisfactory level or above in reading. On the 1998 MSPAP, 41.6% of the state's third graders and 40.4% of the state's fifth graders scored at the satisfactory level or above in reading (MSDE, 1998).

Results from the 1996 NAEP mathematics assessments were not as similar to 1996 MSPAP results. For example, 22% of Maryland fourth graders performed at or above the "proficient" level in the 1996 NAEP mathematics assessment; however, on the 1996 MSPAP, 38.7% of the state's third and 47.8% of the state's fifth graders scored at the satisfactory level or above in mathematics. On the 1998 MSPAP, 41.6% of the state's third graders and 47.9% of the fifth graders scored at the satisfactory level or above in mathematics. A content analysis study funded by National Assessment Governing Board concluded that there are differences between MSPAP and NAEP in grade 8 mathematics in the technical, content, and process dimension. These differences, however, are not sufficient to account for the magnitude of the difference between proficient performance on the MSPAP and NAEP (Kenney & Silver, 1999).

Statistical Test Bias

As a technical term, 'test bias' is not easily defined. A reasonable conceptual approach is to consider a test biased if students of the same degree of attainment in what the test measures receive reliably different scores on the test. A test that fits this definition would

then be biased in favor of those who receive the higher scores and against those who receive the lower scores. The difficulty is, in practice, there is no method available to determine whether or not two different students have the same degree of attainment.

In order to overcome the lack of a 'pure' measure of attainment, overall scores on the test are commonly used as the best available measure in order to evaluate 'bias' at the item level. This approach relies on the assumption that bias, if it exists, is presented in some, as opposed to all, the items on the test. Therefore, to the degree that items are identified as biased, it may be true that the test is biased. However, if no items are identified as biased, then it is a reasonable conclusion that test bias is not a threat to test validity.

Differential item functioning (DIF) procedures examine the possibility that non-essential item characteristics may result in misleading poor performance for minority, female, or other defined groups of students. Although the terms item bias and DIF are used interchangeably, DIF does not necessarily imply unfairness. Evidence of DIF is usually considered as a signal to test developers to examine an item more closely to consider whether or not it is defective before using it again.

Items that are biased against groups of students who take the MSPAP, that is, function differently for different student groups diminish construct validity. A measure of DIF generalized from the Linn-Harnisch procedure (1981) is used to flag differentially functioning items. MSDE has studied items flagged for DIF to inform subsequent assessment task development. MSDE examines performance of African-Americans, Asians, and Hispanics in comparison to Caucasians, and examines the performance of females in comparison with males.

During item calibration, the item parameters estimated for the items assessing a given subject area are used to score **all** of the examinees in the calibration sample. The examinees for each target group (e.g., African American) are then sorted into ten equally numerous score categories (deciles). For each item, using the mean attainment estimate for the examinees of the target group in each decile, the predicted and observed examinee success rates are calculated and compared separately in each decile. A positive difference between the observed and predicted values indicates that the target group members in that decile did better than expected. The positive differences are summed to obtain a positive difference value, D+. Similarly, a negative difference indicates that the target group members in that decile did less well then was expected. The negative differences are also summed to obtain a negative difference value, D-. These two sums of differences are summed to obtain an overall difference, D.

DIF was defined in terms of overall differences in performance and in terms of decile group differences. Items for which |D| was greater or equal to 0.10 were flagged as exhibiting DIF or biased. Items for which |D| was less than 0.10 were called unbiased unless D- were less than or equal to –0.10 or D+ was greater or equal to 0.10. Table 28

presents the number of items for MSPAP 1998 being flagged as exhibiting DIF using the criterion described above. It can be seen that no item was flagged for bias either in favor or against African American target groups in any content area at any grade level. While present, the small numbers of flagged items in the Asian, Hispanic, and female groups may be the result of statistical imprecision due to the relative small sizes of these groups in Maryland.

See Table 28

Consequential Validity Evidence

MSDE, in conjunction with the University of Pittsburgh, is conducting a study to examine the impact of MSPAP on curriculum, instructional and assessment practices, student performance, staff development, and school-based decision-making. It will also examine how the impact varies by content area (reading, writing, language usage, mathematics, science, and social studies), school characteristics (percent minority students, percent free or reduced lunch, MSPAP performance), and grade level (3, 5, 8 and off-grades 2, 4, 7).

Evidence is being collected at system, school, and classroom levels via questionnaires, interviews, and reviews of curriculum, assessment, and professional development materials.

Conclusion

MSPAP scores, in combination with other performance measure, are used to determine school performance consequences such as state mandated intervention in schools failing to demonstrate progress, and rewards for schools consistently making significant improvement.

Validity evidence and other technical information provide reasonably strong assurance that MSPAP scores can be appropriately used for evaluating school performance and guiding school improvement.

Score Interpretation

Two types of scores are available and relevant to school performance and for use in school improvement planning: scale scores and outcome scores. These two types of MSPAP scores are discussed below. For more detailed discussions about score interpretation of MSPAP, consult "Score Intepretation Guide" (MSDE, 1997).

Scale Scores

MSPAP was designed to produce scale scores for the content areas of reading, writing, language usage, mathematics, science, and social studies. MSPAP scale scores indicate a school's level of performance in each content area. MSPAP scale scores range, in general, between 350 and 700 with a mean of approximately 500 and a standard deviation of approximately 50. Scale scores from the same grade level and content area have the same meaning and are directly comparable from year to year. Scale scores are not comparable across grade levels or content areas because of differences in test content and difficulty.

MSPAP scale scores, like other test scale scores, have little intrinsic meaning other than higher scale scores represent higher performance in a content area. Interpretation of the scale scores is aided by proficiency level descriptions. Proficiency level descriptions were developed to help bring meaning to scale scores and to guide interpretation for school performance and improvement.

Proficiency Level Descriptions

The proficiency levels. Proficiency levels and descriptions are intended to inform and guide interpretation of MSPAP scale scores. They describe what students at a particular level generally know and can do in relation to the Maryland Learning Outcomes. The descriptions generally apply to all students at each level rather than to specific students within a level. Individual students whose scale score locates them at a particular proficiency level may or may not be able to demonstrate all of the knowledge, skills, and processes contained in that proficiency level description.

Listed in Appendix C are the scaled score ranges for each proficiency level in each content area and grade. Detailed proficiency descriptions for each content area and grade appear in Appendix B of the Score Intepretation Guide (MSDE, 1997).

As Appendix C indicates, each proficiency level represents a range of performances and of scale scores. For example, grade 3 reading scale scores lower than 490 indicate Level 5 proficiency, those between 490 and 529 indicate Level 4 proficiency, those between 530 and 579 indicate Level 3 proficiency, and so forth.

MSPAP emphasizes high standards of performance. Since MSPAP scale scores can range as low as 350, there is a wide range of scores in Level 5. Generally speaking, students at Level 5 do not consistently demonstrate Level 4 proficiency. However, they may have provided some responses to assessment activities that, with increased consistency, would have placed them at Level 4.

Proficiency level descriptions and proficiency cut scores were established by committees of teachers, principals, content area supervisors, and assistant superintendents. The

committees matched MSPAP items to proficiency level descriptions of Proficiency Levels 1-5, and used the resulting item classifications to establish the location of the cut scores between proficiency levels.

Development of the descriptions. The committee that established the proficiency level cut scores also developed descriptions for each level. For both the establishment and refinement of the descriptions, the committee examined each assessment activity at a proficiency level, the accompanying scoring criteria for each activity, and student responses to each activity. They used their professional judgment to determine and list the knowledge, skills, and processes each activity required of students and to synthesize the lists of required knowledge, skills, and processes into descriptions, in Maryland learning outcomes terms, of what students at each proficiency level know and can do.

Interpretation and use of the proficiency levels and proficiency level descriptions. Proficiency level descriptions apply generally to any group of students, based on performances by all students and schools in Maryland. The descriptions are not customized specifically for individual students, single schools, or other groups.

School Performance Standards

A cornerstone of the Maryland School Performance Program (MSPP) is the process of setting standards of satisfactory and excellent performance levels for schools to meet by 2000.

Development of the standards for MSPAP followed the same procedures used in establishing the school performance standards for all areas reported in the annual *Maryland School Performance Report.* A state Standards Committee researched information on standard setting, identified criteria for standards, and defined the terms *satisfactory* and *excellent.*

> Satisfactory performance denotes a level of performance that is realistic and rigorous for schools, school systems, and the state. It is an acceptable level of performance on a given variable, indicating proficiency in meeting the needs of students.
>
> Excellent performance denotes a level of performance that is highly challenging and clearly exemplary for schools, school systems, and the state. It is a distinguished level of performance on a given variable, indicating outstanding accomplishment in meeting the needs of students (Thorn, Moody, McTighe, Kelly, & Peiffer, 1990, page 7).

Two groups participated in the standards setting process:

> A 20 member Standards Committee of teachers, administrators, content area and assessment specialists, parents, students, university professors, and
>
> A 17 member Standards Council of representatives of local boards of education, teacher's unions, businesses, students, and the Maryland General Assembly.

The process of setting standards included several steps. Initially, the Standards Committee recommended a proficiency level to describe satisfactory and excellent performance and the percentage range of students who should score at these levels (i.e., 60% to 80% at the satisfactory level). These recommendations were reviewed by the Standards Council who refined this work to describe satisfactory and excellent performance by proficiency level and set a percent of students who should be in each category. These two steps depended on a group decision reached though a convergence process.

The recommendations from the Standards Council were reviewed by the State Board of Education and comments were given through public meetings. Following the public meetings, the standards were formally adopted by the State Board of Education.

The Standards Committee recommended level 3 as the proficiency level that describes satisfactory performance and levels 1 and 2 as the proficiency levels that describe excellent performance. Once the ranges for satisfactory and excellent school performance were established, the recommendations were forwarded to the Standards Council. They were asked to choose a single percentage for each standard for school performance. The Council concurred with the definitions for satisfactory and excellent performance. In addition, the Council recommended 70% for satisfactory and 25% for excellent. For a given school to achieve satisfactory performance in a particular area/grade level, 70% of students must achieve satisfactory performance (level 3 and above). To achieve excellent performance, a school must meet the satisfactory requirement and 25% of these students must achieve excellent performance (level 2 and above). The State goal is that all schools will reach the satisfactory standards by the year 2000.

Interpretation and use of school performance standards for school improvement planning. The score reports produced by MSDE for each school system and school contain numbers and percentages of students at each proficiency level and at satisfactory and excellent standards. School and system staff use these percentages, along with the proficiency level descriptions, to evaluate their school's performance in relation to the Maryland Learning Outcomes. They also use this information to assess their school's progress in reaching standards.

Only those students tested are considered when determining a school's proficiency level, because of the focus on the strengths and weaknesses of the students in the school. Since the school performance standards focus on how well a school is performing on the outcomes, any student who should have been tested is included in the calculation[1]. This includes students who were excused from the MSPAP test administration and students who were absent during the test administration. Therefore, proficiency level percentages may be higher than standards percentages, because the proficiency level percentages are usually based on a smaller number of students.

Individual Student Scale Scores

Scale scores and outcome scores for individual students are not interpretable because each student takes only one-third of the total test. Since the primary focus of MSPAP is school performance rather than individual performance, individual student scores are not to be used for decisions for individual student's performance.

Outcome Scores

Within each of the six content areas assessed on MSPAP, i.e., reading, there are more specific outcomes, i.e., reading to be informed. Outcome scores are based on subsets of items that comprise a content area scale. These scores are the scores that would be expected on an outcome if a student had taken all of the items which measure that outcome. For an outcome score to be reported, at least four measures of the outcome must be present in the test form that the student took. There are two types of outcome scores: Outcome Scores and Outcome Scale Scores.

Outcome Scores. MSPAP outcome scores range from 0 to 100% and are reported for each outcome assessed in each MSPAP content area. T hey are conceptually analogous to Maryland Functional Testing Program domain scores and can be interpreted like these scores[2]. Outcome scores indicate the proportion of mastery of the knowledge, skills, processes and other requirements that comprise an outcome area. In other words, the MSPAP school outcome score is the average percentage of all score points available on that outcome that a school achieved across all test clusters administered in the school.

Outcome scores are not directly comparable across grades and content areas within a grade, nor are they directly comparable across years because of differences in content and test difficulty. However, they can be compared using information on the relative difficulty of each outcome. Moreover, outcome scores cannot be directly linked to MSPAP proficiency levels.

Interpretation and Use of Outcome Scores. School improvement teams use profiles of a school's Outcome Scores in a content area along with other information about a school, to

determine a school's instructional program's relative strengths and weaknesses in each MSPAP content area.

Content area relative difficulty values are reported on Table 29. Relative difficulty refers to the average proportion of the maximum possible score for an outcome across clusters. The relative outcome difficulty index ranges from 0 to 100%. Lower percentages indicate harder outcomes, and conversely, higher percentages indicate easier outcomes. This information is used in conjunction with outcome score averages. An index of relative difficulty was developed because of the desire to compare outcome score averages within each content area to one another.

See Table 29

Outcome Scale Scores. Outcome scale scores are directly comparable across outcomes in the same content area, across years, and to the MSPAP proficiency levels. These scores are expressed on the MSPAP scale score scale and range, as are the content area scale scores, from 350 to 700. Therefore, they can be interpreted in relationship to the underlying score scale and proficiency levels.

MSPAP Score Reports

The four main types of MSPAP score reports are: Maryland School Performance Standards Reports, Proficiency Level and Participation Reports, Outcome Score Reports, and Outcome Scale Score Reports. MSDE provides these reports at the state, school system, and school levels.

MSPAP Standards Reports. These reports provide information on the percentages of students at satisfactory and excellent levels of performance and indicate whether the standards for satisfactory and excellent school performance have been met. Information on the numbers and percentages of students by grade, content area, race, and gender is available in the MSPAP Disaggregated Standards Report.

MSPAP Proficiency Level and Participation Reports. These reports provide the numbers and percentages of test takers at each of the five MSPAP proficiency levels. They also report numbers and percentages of students who completed assessment activities in each MSPAP content area and received a scale score. Also, numbers and percentages of students who were absent, excused, or exempted from the MSPAP test administration are reported. Information on the numbers and percentages of students by grade, content area, race, and gender is available in the Disaggregated Proficiency Level and Participation Report.

MSPAP Outcome Score Reports. Outcome Score Reports contain the average outcome score, or percentage of mastery of an outcome, for a school, school system, or the state. The Outcome Score Reports also include percentages of students in four outcome score ranges: 0-25, 26-50, 51-75, and 76-100. This information is intended to provide a general idea of the percentage of students who have displayed little or no mastery of the knowledge, skills, and processes required in an outcome (i.e., those in the outcome score range 0-25) and the percentage who have displayed near complete mastery of the outcome (i.e., those in the range 76-100).

MSPAP Outcome Scale Score Reports. The Outcome Scale Score Reports contain the median outcome scale score for each learning outcome. The median (50th percentile), the interquartile range (25th to 75th percentiles) and the 5th to 95th. Outcome Scale Score Reports can be used to compare outcome performance within a content area. Unlike outcome scores, outcome scale scores can be compared in a content area because the outcome scale scores have been adjusted for difficulty.

It is important not to over interpret the relationship between outcome scale scores and proficiency levels. Outcome scale scores represent performance on activities that measure only that outcome. In contrast, proficiency levels are established based on all the outcomes in a content area.

References

Allen, M., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) New York: American Council on Education.

Binkley M., Atash, M. N., & Bourque, M. (in press). Standard setting and reporting. In T. Husen and N. Postlethwaite (Eds.), *The International Encyclopedia of Education*, 2nd ed.

Burket, G. R. (1991). *PARDUX, Version 1.4.* Monterey CA: CTB Macmillan/McGraw Hill.

Burket, G. R. (1991). *FLUX Version 1.0.* Monterey, CA: CTB Macmillan/McGraw-Hill.

CTB Macmillan/ McGraw Hill. (1992). *Final technical report: Maryland School Performance Assessment Program, 1991.* (Available from the Maryland State Department of Education, Baltimore, MD.)

Ebel, R. L. (1979). *Essentials of educational measurement*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Kenney, P, & Sliver, E. (1998). Content Analysis Project – State and NAEP Mathematics Assessment. *Report of Results from the Maryland MAEP Study*. Learning Research and Development Center, University of Pittsburg.

Kentucky Department of Education. (1997). *KIRIS Accountability Cycle I Technical Manual*: Lexington: Author.

Linn, R. L. & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Mamillan/McGraw-Hill School Publishing Company (1993). *Reflecting Diversity- Multicultural guidelines for educational publishing professionals*. New York: Author.

Maryland State Department of Education (1993a). *Scoring MSPAP: A Teacher's Guide*. Baltimore: Author.

Maryland State Department of Education. (1993b). *Technical report: 1992 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1994). *Technical report: 1993 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1995). *Technical report: 1994 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1996). *Technical report: 1995 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1997). *Technical report: 1996 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1997). *Test administration and coordination manual, 1997*. Baltimore: Author.

Maryland State Department of Education. (1997). *Score Intepretation Guide, Maryland School Performance Assessment Program - 1997 MSPAP and Beyond, 1997*. Baltimore: Author.

Maryland State Department of Education. (1998). *Maryland School Performance Report, 1998*. Baltimore: Author.

Measurement Incorporated. (1998). 1998 *Maryland School Performance Assessment Program scoring report*. (Available from the Maryland State Department of Education, Baltimore, MD)

National Evaluation System Inc. (1991*). Bias Issues in Test Development.* Amerst, MA: Author.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) New York: American Council on Education/ Macmillan.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Thorn, P., Moody, M., McTighe, J., Kelly, N., & Peiffer, R. (1990, April). *Establishing standards for Maryland's School Systems: A systemic approach*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.

Westat, Inc. (1998). *1998 MSPAP Field Test Report*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.

Westat, Inc. (1994). *Establishing proficiency levels and descriptions for the 1994 MSPAP assessment program.* (Available from the Maryland State Department of Education, Baltimore, MD)

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

TABLES

TABLE 1

Numbers of Teams, Readers, and Scoring Leaders by Site, Cluster, and Grade

Site	Grade/ Cluster	Number Of Teams	Target Number of Readers	Number Of Coordinators	Number Of Leaders
Western Tech	3B	4	68	4	4
	5B	4	83	4	4
	8B	4	83	4	4
Total		12	234	12	12
Chesapeake	3C	4	78	4	4
	5C	4	82	4	4
	8C	4	82	4	4
Total		12	242	12	12
Waldorf	3A	4	91	4	4
	8A	4	95	4	4
Total		8	186	8	8
Grasonville	5A	4	82	4	4
GRAND TOTAL		36	744	36	36

TABLE 2
1998 READER ACCURACY SET MEAN SCORES BY TEAM
GRADE 3

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	AVERAGE
A1	81	81	87	84	80	84	83
A2	91	89	87	97	89	85	90
A3	79	92	76	83	79	90	83
A4	98	79	97	74	100	88	89
B1	98	79	97	74	100	66	89
B2	86	88	92	95	86	86	89
B3	93	90	98	85	85	75	88
B4	76	86	75	76	77	82	79
C1	82	86	86	86	91	80	85
C2	84	80	81	--	--	--	82
C3	98	94	98	96	97	98	97
C4	92	90	93	91	96	92	92

43

TABLE 3
1998 READER ACCURACY SET MEAN SCORES BY TEAM
GRADE 5

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	AVERAGE
A1	66	74	75	78	61	74	71
A2	87	76	82	79	81	89	82
A3	93	91	90	92	--	--	92
A4	82	86	84	85	100	88	88
B1	91	91	90	85	88	89	89
B2	91	92	89	93	95	95	93
B3	79	80	82	85	84	81	82
B4	81	82	88	74	80	84	82
C1	91	93	88	92	90	91	91
C2	73	81	87	--	--	--	80
C3	98	96	99	96	94	--	97
C4	81	92	89	92	86	92	89

TABLE 4
1998 READER ACCURACY SET MEAN SCORES BY TEAM
GRADE 8

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	AVERAGE
A1	96	90	89	91	96	--	92
A2	83	82	83	85	83	83	83
A3	73	74	69	75	73	78	74
A4	86	86	85	82	86	88	86
B1	93	91	90	93	96	93	93
B2	80	87	72	70	73	78	77
B3	85	85	92	88	82	88	87
B4	84	75	81	89	82	86	83
C1	78	85	94	89	97	95	90
C2	90	90	80	94	86	88	90
C3	78	90	88	93	87	91	88
C4	79	75	88	82	86	83	82

TABLE 5
1998 FREQUENCY OF ACCURACY SET MEAN SCORES BY GRADE

Grade	Less than 70 Percent	70-79 Percent	80-89 Percent	90-100 Percent
3	0(0%)	12 (17%)	29 (42%)	28 (41%)
5	2(3%)	9 (14%)	29 (44%)	26 (39%)
8	1 (1%)	14 (20%)	35 (49%)	21 (30%)
ALL GRADES	3 (1%)	35 (17%)	93 (45%)	75 (36%)

TABLE 6
1998 READER ACCURACY SET MEAN SCORES BY CONTENT AREA
GRADE 8

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	AVERAGE
Matematics								
A1	96	90	89	91	96	--	92	89
B1	93	91	90	93	96	93	93	94
C1	78	85	94	89	97	95	90	93
Social studies								
A2	83	82	83	85	83	83	83	78
B2	80	87	72	70	73	78	77	74
C2	90	90	90	94	86	88	90	87
Science								
A3	73	74	69	75	73	78	74	84
B3	85	85	92	88	82	88	87	87
C3	78	90	88	93	87	91	88	87
Writing								
A4	86	86	85	82	86	88	86	86
B4	84	75	81	89	82	86	83	89
C4	79	75	88	82	86	83	82	86

*Note: Content areas are somewhat integrated.

TABLE 7
1998 FREQUENCY OF ACCURACY SET MEAN SCORES BY CONTENT AREA
GRADE 8

Content	Less than 70 Percent	70-79 Percent	80-89 Percent	90-100 Percent
Mathematics	0 (0%)	1 (6%)	3 (18%)	13 (76%)
Social Studies	0 (0%)	4 (22%)	10 (56%)	4 (22%)
Science	1 (6%)	6 (33%)	7 (39%)	4 (22%)
Writing	0 (0%)	3 (17%)	15 (83%)	0 (0%)
All Content Areas	1 (1%)	11 (15%)	41 (56%)	20 (27%)

TABLE 8. Summary Findings from Calibrations

Content Area/ Cluster	Sample Size	No. of Items[1]	No. of Items Deleted[2] GA	MSDE	Fit	No. Items with Hand-Estimated Parameters	No. of Items with Fit > Criterion[3]	No. of Students at Min./Max.
Reading								
3A*	7,499*	24*	0	0	0	0	0	119
3B	7,499	12	0	0	0	0	1	175
3C	7,499	13	0	0	0	0	1	136
5A	7,500	12	0	0	0	0	1	159
5B*	7,500	30*	0	0	0	0	2	64
5C	7,500	12	0	0	0	0	1	111
8A	7,501	13	0	0	0	0	0	208
8B	7,501	13	0	0	0	0	1	254
8C*	7,501	33*	0	0	0	0	9	84
Writing/Language Usage								
3A*	7,499*	19*	0	0	0	1	2	779
3B	7,499	12	0	0	0	0	2	787
3C	7,499	12	0	0	0	0	0	759
5A	7,500	11	0	0	0	0	3	424
5B*	7,500	21*	0	0	0	0	6	359
5C	7,500	11	0	0	0	0	0	710
8A	7,501	13	0	0	0	1	4	669
8B	7,501	11	0	0	0	0	7	449
8C*	7,501	23*	0	0	0	1	8	256
Math Content								
3A	7,499	23	0	0	0	0	0	43
3B	7,499	15	0	0	0	0	2	124
3C	7,499	23	0	0	0	0	0	74
5A	7,500	18	0	0	0	0	2	95
5B	7,500	18	0	0	0	0	1	149
5C	7,500	27	0	0	0	0	0	77
8A	7,501	17	0	0	0	0	0	654
8B	7,501	18	0	0	0	0	4	289
8C	7,501	13	0	0	0	0	3	665

(table 8 continue)

Content Area/ Cluster	Sample Size	No. of Items[1]	No. of Items Deleted[2] GA	MSDE	Fit	No. Items with Hand-Estimated Parameters	No. of Items with Fit > Criterion[3]	No. of Students at Min./Max.
Math Process								
3A	7,499	12	0	0	0	0	1	418
3B	7,499	7	0	0	0	0	7	646
3C	7,499	11	0	0	0	0	2	655
5A	7,500	9	0	0	0	0	5	311
5B	7,500	10	0	0	0	0	0	145
5C	7,500	12	0	0	0	0	2	108
8A	7,502	7	0	0	0	0	4	945
8B	7,502	11	0	0	0	0	2	489
8C	7,502	6	0	0	0	0	5	1484
Science								
3A	7,499	17	0	0	0	0	1	174
3B	7,499	18	0	0	0	0	1	178
3C	7,499	18	0	0	0	0	0	153
5A	7,500	14	0	0	0	0	0	293
5B	7,500	19	0	0	0	0	0	129
5C	7,500	19	0	0	0	0	0	62
8A	7,502	23	0	0	0	0	0	261
8B	7,502	27	0	0	0	0	0	189
8C	7,502	19	0	0	0	0	0	211
Social Studies								
3A	7,499	17	0	0	0	0	0	320
3B	7,499	16	0	0	0	0	0	421
3C	7,499	16	0	0	0	0	1	163
5A	7,500	16	0	0	0	0	0	133
5B	7,500	18	0	0	0	0	0	88
5C	7,500	17	0	0	0	0	0	182
8A	7,502	17	0	0	0	0	0	270
8B	7,502	22	0	0	0	0	0	255
8C	7,502	18	0	0	0	0	0	240

(table 8 continue)

[1] No. of items refers to the number of items defined as assessing each content area prior to scaling and before items were deleted for the reasons specified in the next column. For the Reading and Writing/Language Usage items in 3A, 5B, and 8C, the No. of items is the total number of items in all choice sets; students administered these clusters actually responded to fewer items than the total given.

[2] The reasons for the item deletion are designated as GA signifying group-administration; MSDE signifying a deletion requested by MSDE; and Fit signifying poor fit.

[3] The cut-off Z values used for various N counts are as follows:

N	Z >
1,500	4
2,000	5
3,000	8
4,000	11
5,000	13
6,000	16
7,000	19

* This is a choice cluster. Sample size, the numbers of items, and the number of misfitting items for this cluster varied over the choice sets.

TABLE 9. Detailed Findings from Calibration for Clusters with Choice Sets in Reading and Writing

Content	Cluster	Choice	Sample Size	Number of Items	Number of Items with Fit Exceeding Criterion[1]
Reading					
	3A	Non-choice	7499	6	0
		Choice A	740	6	0
		Choice B	4599	6	0
		Choice C	2160	6	0
	5B	Non-choice	7500	6	0
		Choice A	2248	6	0
		Choice B	1057	6	0
		Choice C	2226	6	0
		Choice D	1959	6	0
	8C	Non-choice	7502	5	0
		Choice A	1156	7	3
		Choice B	4707	7	3
		Choice C	1187	7	0
		Choice D	452	7	1
Writing					
	3A	Non-choice	7499	2	2
		Story	4619	1	0
		Poem	2443	1	0
		Play	437	1	0
	5B	Non-choice	7500	2	1
		Story	4049	1	0
		Poem	2992	1	0
		Play	459	1	0
	8C	Non-choice	7502	2	1
		Story	3883	1	0
		Poem	2753	1	0
		Play	302	1	0
		Other	564	1	0

1 See footnote of Table 8 for the fitting criterion

TABLE 10. Cluster Equating Results

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
Reading				
3A	400	650	5	1
3B(T)*	400	650	6	1
3C	400	650	5	0
5A	375	675	2	1
5B	375	675	4	0
5C(T)*	375	675	2	0
8A	375	650	4	0
8B (T)*	375	650	4	1
8C	375	650	1	1
Writing				
3A	455	635	22	1
3B	455	635	27	1
3C(T)*	455	635	26	3
5A	440	595	19	5
5B	440	595	18	8
5C(T)*	440	595	18	8
8A(T)*	425	625	25	8
8B	425	625	29	5
8C	425	625	22	6

(table 10 continue)

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
Language Usage				
3A	450	625	10	1
3B(T)*	450	625	11	1
3C	450	625	12	1
5A	425	625	13	2
5B	425	625	12	3
5C(T)*	425	625	14	2
8A	425	625	11	3
8B(T)*	425	625	11	4
8C	425	625	11	3
Math Content				
3A(T)*	375	650	3	0
3B	375	650	2	0
3C	375	650	3	0
5A	400	650	6	0
5B	400	650	6	0
5C(T)*	400	650	6	0
8A(T)*	400	650	11	0
8B	400	650	7	0
8C	400	650	8	1
Math Process				
3A(T)*	375	650	6	0
3B	375	650	8	0
3C	375	650	9	0
5A	400	650	7	1
5B	400	650	7	0
5C(T)*	400	650	7	1
8A	400	650	12	1
8B(T)*	400	650	10	0
8C	400	650	19	1

(table 10 continue)

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
Social Studies				
3A(T)*	400	625	8	0
3B	400	625	9	0
3C	400	625	8	0
5A	400	625	6	0
5B(T)*	400	625	7	0
5C	400	625	6	0
8A	375	650	5	0
8B(T)*	375	650	5	0
8C	375	650	4	0
Science				
3A	375	650	4	0
3B(T)*	375	650	4	0
3C	375	650	4	0
5A	375	650	5	0
5B(T)*	375	650	3	0
5C	375	650	4	0
8A	375	650	4	0
8B(T)*	375	650	3	0
8C	375	650	4	0

TABLE 11
Rater Year Effects Study Performance ($97SS_{97}$) of State Sample on 1997 MSPAP

		State[1]			Sample		
Grade	Scale	Mean	SD	N	Mean	SD	N
3	Reading	513.0	46.8	19,451	514.5	45.8	1,454
	Writing	523.3	48.7	19,921	526.2	48.3	1,454
	Language Usage	524.0	58.5	20,037	525.4	57.6	1,454
	Math Content	517.0	57.2	19,591	518.6	56.7	1,454
	Math Process	518.2	43.1	19,591	520.1	42.5	1,454
	Social Studies	503.4	48.0	19,856	506.1	46.6	1,454
	Science	508.8	53.3	19,635	510.4	52.6	1,454
5	Reading	513.1	46.9	19,161	518.1	43.1	1,433
	Writing	506.5	56.1	19,349	510.2	54.2	1,433
	Language Usage	524.1	58.8	19,357	530.5	55.0	1,433
	Math Content	519.5	53.5	19,201	524.0	50.7	1,433
	Math Process	512.9	55.9	19,201	517.7	52.8	1,433
	Social Studies	517.1	55.5	19,392	519.2	53.6	1,433
	Science	514.8	56.6	19,017	519.6	53.0	1,433
8	Reading	510.2	37.6	17,459	513.5	36.3	1,505
	Writing	502.6	53.0	17,737	508.3	52.2	1,505
	Language Usage	509.1	57.8	17,900	514.4	56.6	1,505
	Math Content	521.0	47.3	18,169	524.8	45.9	1,505
	Math Process	512.8	58.1	18,169	518.1	56.2	1,505
	Social Studies	516.4	53.6	17,961	521.3	50.6	1,505
	Science	523.7	54.3	17,459	528.3	52.4	1,505

[1] State performance results were drawn from the Forms Effect Study carried out for the 1997 MSPAP. The values reported refer to performance on Clusters 3C, 5A, and 8A.

TABLE 12
Rater Year Effects Study Raw Score Comparisons

			Raters Used				
			1997		1998		Mean Diff.
Grade	Scale	N	Mean	SD	Mean	SD	(98 – 97)
3	Reading	1454	9.63	5.25	9.82	5.17	0.19
	Writing	1454	3.03	1.86	2.86	1.85	-0.17
	Language Usage	1454	7.02	5.21	6.10	5.18	-0.92
	Math Content	1454	14.71	6.45	14.84	6.53	0.13
	Math Process	1454	6.69	3.83	6.87	3.95	0.18
	Social Studies	1454	16.17	7.50	16.04	7.57	-0.13
	Science	1454	14.67	6.22	14.82	6.20	0.15
5	Reading	1433	13.44	4.99	13.97	4.82	0.53
	Writing	1433	3.02	1.66	3.00	1.67	-0.02
	Language Usage	1433	6.44	4.80	6.78	4.70	0.34
	Math Content	1433	10.49	4.48	10.81	4.51	0.32
	Math Process	1433	7.08	3.50	7.50	3.67	0.42
	Social Studies	1433	12.83	6.47	14.73	6.89	1.90
	Science	1433	11.65	6.27	11.21	5.98	-0.44
8	Reading	1505	12.36	5.01	12.11	4.98	-0.25
	Writing	1505	3.47	1.92	3.57	1.93	0.10
	Language Usage	1505	7.53	4.79	8.07	4.76	0.54
	Math Content	1505	13.65	8.66	13.68	8.63	0.03
	Math Process	1505	5.02	3.47	5.17	3.53	0.15
	Social Studies	1505	13.11	6.25	12.92	6.27	-0.19
	Science	1505	11.05	6.93	10.61	6.69	-0.44

TABLE 13
1992, 1993, 1994, 1995, 1996, 1997, and 1998 Rater Year Effects Studies:
Comparison of Results in Terms of Standardized Raw Score Mean Differences[1]

		Rater Effects Study						
Grade	Scale	1992	1993	1994	1995	1996	1997	1998
3	Reading	0.0	-0.2	0.0	0.0	0.0	-0.1	0.0
	Writing	-0.2	-0.2	0.0	0.2	0.0	0.0	-0.1
	Language Usage	-0.2	-0.4	0.0	0.2	0.0	-0.1	-0.1
	Math Content	0.1	0.0	-0.1	0.1	0.0	0.0	0.0
	Math Process	0.2	0.0	0.0	0.1	0.0	-0.1	0.0
	Social Studies[2]	---	-0.6	-0.1	0.1	0.1	0.0	0.0
	Science[2]	---	-0.2	0.1	0.0	0.0	0.0	0.0
5	Reading	0.3	-0.2	0.1	0.1	-0.1	0.3	0.1
	Writing	0.4	-0.1	0.0	0.1	0.0	-0.1	0.0
	Language Usage	0.3	-0.2	0.0	-0.1	0.0	0.0	0.1
	Math Content	0.1	-0.1	0.1	0.0	0.0	0.0	0.1
	Math Process	0.2	0.0	0.1	0.0	0.0	0.0	0.1
	Social Studies[2]	---	0.1	0.2	0.0	0.1	0.1	0.2
	Science[2]	---	-0.1	0.1	0.1	0.0	0.2	-0.1
8	Reading	0.0	0.1	-0.1	-0.2	-0.1	0.1	0.0
	Writing	0.0	0.0	0.2	-0.1	0.1	0.1	0.0
	Language Usage	-0.2	0.1	-0.1	0.0	0.1	0.3	0.1
	Math Content	0.1	-0.1	0.0	-0.1	0.0	0.0	0.0
	Math Process	0.1	-0.1	-0.1	-0.1	-0.1	0.0	0.0
	Social Studies[2]	---	-0.1	-0.1	0.0	-0.1	0.0	0.0
	Science[2]	---	-0.2	0.0	-0.2	0.0	0.1	0.0

[1] These differences were obtained by dividing the difference between the current and prior year mean ratings by the square root of the pooled variances of these ratings.

[2] This subject was not assessed in this grade in 1991, so comparisons involving 1991 ratings are not available.

TABLE 14
Rater Year Effects Study Transformation Values

Grade	Scale	Multiplier R_1	Addend R_2	(A) $(R_1*500)+R_2$	(A) - 500[1]
3	Reading	1.034	-18.752	498.248	-2
	Writing	1.008	0.666	504.666	5
	Language Usage	0.924	54.089	515.984	16
	Math Content	0.984	7.590	499.590	0
	Math Process	0.973	12.315	498.815	-1
	Social Studies	0.991	7.093	502.593	3
	Science	1.008	-1.816	502.184	2
5	Reading	1.027	-19.964	493.536	-6
	Writing	0.900	51.809	502.019	2
	Language Usage	1.045	-27.736	494.764	-5
	Math Content	1.000	-3.000	497.000	-3
	Math Process	0.926	32.850	495.850	-4
	Social Studies	0.958	8.072	487.222	-13
	Science	1.020	-8.432	501.568	2
8	Reading	1.027	-10.841	502.659	3
	Writing	0.978	4.997	494.197	-6
	Language Usage	1.007	-11.329	492.171	-8
	Math Content	0.990	5.736	500.736	1
	Math Process	0.996	0.761	498.761	- 1
	Social Studies	1.000	1.000	501.000	1
	Science	1.020	-7.914	502.086	2

[1] Numbers in this column were purposely rounded to improve their comprehensibility

TABLE 15
Performance of State on 1997 MSPAP and 1998 Equating Sample on 1997 MSPAP

		State[1] ($97SS_{97}$)			Sample ($97SS_{98}$)		
Grade	Scale	Mean	SD	N	Mean	SD	N
3	Reading	513.0	46.8	19,451	516.6	48.0	2,323
	Writing	523.3	48.7	19,921	525.4	47.9	2,323
	Language Usage	524.0	58.5	20,037	525.2	57.9	2,323
	Math Content	517.0	57.2	19,591	516.0	57.2	2,323
	Math Process	518.2	43.1	19,591	518.4	43.8	2,323
	Social Studies	503.4	48.0	19,856	509.2	48.5	2,323
	Science	508.8	53.3	19,635	509.2	53.8	2,323
5	Reading	513.1	46.9	19,161	520.7	47.7	2,454
	Writing	506.5	56.1	19,349	512.8	53.7	2,454
	Language Usage	524.1	58.8	19,357	532.4	56.2	2,454
	Math Content	519.5	53.5	19,201	524.0	56.2	2,454
	Math Process	512.9	55.9	19,201	512.7	56.8	2,454
	Social Studies	517.1	55.5	19,392	520.2	54.1	2,454
	Science	514.8	56.6	19,017	526.0	53.1	2,454
8	Reading	510.2	37.6	17,459	509.2	40.7	2,389
	Writing	502.6	53.0	17,737	503.9	53.3	2,389
	Language Usage	509.1	57.8	17,900	514.1	59.4	2,389
	Math Content	521.0	47.3	18,169	523.7	47.1	2,389
	Math Process	512.8	58.1	18,169	515.6	59.4	2,389
	Social Studies	516.4	53.6	17,961	517.7	53.9	2,389
	Science	523.7	54.3	17,459	525.2	56.4	2,389

[1] State performance results were drawn from the Forms Effect Study carried out for the 1997 MSPAP. The values reported refer to performance on Clusters 3C, 5A, and 8A.

TABLE 16
Equating Study Transformation Values

Grade	Scale	Multiplier T_1	Addend T_2	(A) $(T_1*500)+T_2$	(A) - 500[1]
3	Reading	0.808	116.710	520.710	21
	Writing	0.817	116.584	525.084	25
	Language Usage	1.409	-181.871	522.629	23
	Math Content	1.093	-27.383	519.117	19
	Math Process	0.610	218.543	523.543	24
	Social Studies	0.929	49.439	513.939	14
	Science	1.039	-6.838	512.662	13
5	Reading	0.875	80.645	518.145	18
	Writing	1.199	-87.518	511.982	12
	Language Usage	1.071	0.731	536.231	36
	Math Content	1.089	-17.370	527.130	27
	Math Process	0.973	28.859	515.359	15
	Social Studies	1.093	-25.230	521.270	21
	Science	1.010	23.112	528.112	28
8	Reading	0.690	162.113	507.113	7
	Writing	0.989	8.476	502.976	3
	Language Usage	1.181	-82.285	508.215	8
	Math Content	0.805	127.133	529.633	30
	Math Process	1.085	-23.308	519.192	19
	Social Studies	1.017	12.943	521.443	21
	Science	1.039	9.252	528.752	29

[1] Numbers in this column were purposely rounded to improve their comprehensibility.

TABLE 17
Comparison of 1997 and 1998 MSPAP Performance by Grade and Scale

Grade	Scale	1997 State Means	1998 State Means	98 - 97 Difference
3	Reading	513.9	519.7	5.8
	Writing	521.6	523.5	1.9
	Language Usage	524.3	524.0	-0.3
	Math Content	516.1	517.2	1.1
	Math Process	516.9	514.0	-2.9
	Total Math	516.8	515.9	-0.9
	Social Studies	503.1	509.0	5.9
	Science	508.6	509.4	0.8
5	Reading	513.7	516.0	2.3
	Writing	506.9	508.0	1.1
	Language Usage	523.4	529.7	6.3
	Math Content	518.5	520.2	1.7
	Math Process	511.7	511.4	-0.3
	Total Math	515.5	516.2	0.7
	Social Studies	516.7	516.5	-0.2
	Science	514.7	521.3	6.6
8	Reading	510.9	508.1	-2.8
	Writing	502.8	503.9	1.1
	Language Usage	510.2	510.3	0.1
	Math Content	521.0	523.7	2.7
	Math Process	513.0	513.7	0.7
	Total Math	517.2	519.1	1.9
	Social Studies	516.9	518.0	1.1
	Science	525.2	528.8	3.6

TABLE 18. Coefficient Alpha for 1998 MSPAP Content Areas

Grade 3

	Cluster A	Cluster B	Cluster C
Reading	.83	.84	.82
Writing	.65	.74	.78
Language Usage	.91	.93	.93
Math Total	.86	.84	.88
Math Content	.85	.82	.85
Math Process	.75	.71	.75
Science	.84	.81	.87
Social Studies	.86	.86	.80

Grade 5

	Cluster A	Cluster B	Cluster C
Reading	.86	.83	.81
Writing	.69	.65	.71
Language Usage	.90	.91	.91
Math Total	.82	.88	.90
Math Content	.81	.86	.89
Math Process	.70	.75	.74
Science	.81	.83	.79
Social Studies	.85	.82	.86

Grade 8

	Cluster A	Cluster B	Cluster C
Reading	.86	.88	.87
Writing	.80	.70	.70
Language Usage	.94	.92	.92
Math Total	.90	.88	.86
Math Content	.88	.87	.84
Math Process	.71	.82	.71
Science	.86	.89	.83
Social Studies	.89	.90	.89

Note: Clusters 3A, 5B, and 8C are choice clusters.
The reported alpha for the choice cluster are the average alpha across all choices.

TABLE 19. Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 3

		Cluster		
Reading	**Scale Score**	**3A**	**3B**	**3C**
SE at HOSS	650	44	41	41
SE at Level 1/2	620	29	26	30
SE at Level 2/3	580	20	17	20
SE at Level 3/4	530	15	15	16
SE at Level 4/5	490	16	17	16
SE at LOSS	400	42	37	36
Writing				
SE at HOSS	635	35	38	42
SE at Level 1/2	614	29	35	33
SE at Level 2/3	577	25	26	25
SE at Level 3/4	528	31	23	22
SE at LOSS	455	62	42	39
Language Usage				
SE at HOSS	625	21	23	23
SE at Level 1/2	620	20	22	22
SE at Level 2/3	576	18	16	17
SE at Level 3/4	521	19	16	18
SE at LOSS	450	31	32	29
Math Content				
SE at HOSS	650	28	52	32
SE at Level 1/2	626	25	37	26
SE at Level 2/3	583	20	29	20
SE at Level 3/4	531	18	21	19
SE at Level 4/5	489	20	20	21
SE at LOSS	375	46	44	41
Math Process				
SE at HOSS	650	33	67	32
SE at Level 1/2	626	21	37	23
SE at Level 2/3	583	14	23	16
SE at Level 3/4	531	14	14	15
SE at Level 4/5	489	25	26	20
SE at LOSS	375	111	124	124
Science				
SE at HOSS	650	32	31	34
SE at Level 1/2	619	26	24	26
SE at Level 2/3	580	20	19	20
SE at Level 3/4	527	19	16	17
SE at Level 4/5	488	20	17	18
SE at LOSS	375	41	42	37
Social Studies				
SE at HOSS	625	23	26	30
SE at Level 1/2	622	21	26	30
SE at Level 2/3	580	16	16	20
SE at Level 3/4	525	15	15	18
SE at Level 4/5	495	17	17	20
SE at LOSS	400	44	45	38

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

TABLE 20. Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 5

		Cluster		
Reading	**Scale Score**	**5A**	**5B**	**5C**
SE at HOSS	675	57	41	62
SE at Level 1/2	620	28	24	38
SE at Level 2/3	580	21	17	24
SE at Level 3/4	530	15	15	19
SE at Level 4/5	490	15	17	18
SE at LOSS	375	45	46	34
Writing				
SE at HOSS	595	42	48	36
SE at Level 2/3	567	39	42	34
SE at Level 3/4	522	38	38	35
SE at Level 4/5	488	39	48	39
SE at LOSS	440	43	47	48
Language Usage				
SE at HOSS	625	21	26	20
SE at Level 1/2	597	16	17	16
SE at Level 2/3	567	14	14	14
SE at Level 3/4	533	15	14	15
SE at LOSS	425	62	48	68
Math Content				
SE at HOSS	650	37	31	27
SE at Level 1/2	617	31	22	19
SE at Level 2/3	575	25	19	16
SE at Level 3/4	520	21	18	16
SE at Level 4/5	473	22	20	21
SE at LOSS	400	37	42	36
Math Process				
SE at HOSS	650	59	39	43
SE at Level 1/2	617	47	29	28
SE at Level 2/3	575	34	23	24
SE at Level 3/4	520	27	24	23
SE at Level 4/5	473	26	29	27
SE at LOSS	400	44	46	44
Science				
SE at HOSS	650	28	25	27
SE at Level 1/2	625	22	22	23
SE at Level 2/3	580	17	18	20
SE at Level 3/4	525	18	19	22
SE at Level 4/5	484	24	21	24
SE at LOSS	375	77	49	46
Social Studies				
SE at HOSS	625	25	29	20
SE at Level 1/2	619	24	28	19
SE at Level 2/3	580	20	23	18
SE at Level 3/4	529	19	33	19
SE at LOSS	400	38	39	43

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

TABLE 21. Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 8

		Cluster		
Reading	**Scale Score**	**8A**	**8B**	**8C**
SE at HOSS	650	50	73	76
SE at Level 1/2	650	50	73	76
SE at Level 2/3	580	18	20	26
SE at Level 3/4	530	12	12	11
SE at Level 4/5	490	13	11	10
SE at LOSS	375	47	54	47
Writing				
SE at HOSS	625	64	57	67
SE at Level 2/3	551	29	31	37
SE at Level 3/4	505	25	29	28
SE at LOSS	425	35	41	33
Language Usage				
SE at HOSS	625	26	29	38
SE at Level 2/3	565	14	18	19
SE at Level 3/4	509	14	20	17
SE at Level 4/5	474	15	18	17
SE at LOSS	425	27	22	22
Math Content				
SE at HOSS	650	37	24	43
SE at Level 1/2	618	19	17	22
SE at Level 2/3	579	12	12	13
SE at Level 3/4	525	12	15	14
SE at Level 4/5	481	23	19	14
SE at LOSS	400	93	50	88
Math Process				
SE at HOSS	650	56	28	43
SE at Level 1/2	618	34	21	26
SE at Level 2/3	579	25	18	22
SE at Level 3/4	525	26	22	28
SE at Level 4/5	481	32	28	40
SE at LOSS	400	76	54	102
Science				
SE at HOSS	650	21	28	26
SE at Level 1/2	619	16	22	19
SE at Level 2/3	576	14	16	17
SE at Level 3/4	532	14	14	17
SE at Level 4/5	482	19	14	23
SE at LOSS	375	54	39	58
Social Studies				
SE at HOSS	650	31	30	38
SE at Level 1/2	620	24	22	31
SE at Level 2/3	582	17	17	20
SE at Level 3/4	530	15	13	15
SE at Level 4/5	495	16	14	15
SE at LOSS	375	46	46	44

Note: HOSS is the highest obtainable scale score. LOSS is the lowest obtainable scale score.

TABLE 22. Between Content Area Scale Score Correlations for Grade 3

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.61	1.00				
Lang. Usage	.62	.77	1.00			
Mathematics	.69	.59	.61	1.00		
Science	.79	.64	.65	.76	1.00	
Social Studies	.73	.62	.64	.71	.78	1.00

Note: N ranges from 57,980 to 63,337.

TABLE 23. Between Content Area Scale Score Correlations for Grade 5

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.58	1.00				
Lang. Usage	.61	.76	1.00			
Mathematics	.62	.59	.63	1.00		
Science	.68	.57	.62	.73	1.00	
Social Studies	.68	.60	.63	.69	.71	1.00

Note: N ranges from 55,882 to 61,357.

TABLE 24. Between Content Area Scale Score Correlations for Grade 8

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.67	1.00				
Lang. Usage	.70	.82	1.00			
Mathematics	.62	.60	.64	1.00		
Science	.75	.63	.66	.75	1.00	
Social Studies	.69	.64	.66	.70	.77	1.00

Note: N ranges from 52,348 to 55,841.

TABLE 25. Between Content Area Scale Score Correlations at School Level for Grade 3

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.94	1.00				
Lang. Usage	.91	.95	1.00			
Mathematics	.94	.91	.88	1.00		
Science	.96	.94	.90	.96	1.00	
Social Studies	.96	.93	.90	.96	.98	1.00

Note: N=808

TABLE 26. Between Content Area Scale Score Correlations At School Level for Grade 5

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.91	1.00				
Lang. Usage	.91	.95	1.00			
Mathematics	.90	.91	.90	1.00		
Science	.94	.92	.91	.96	1.00	
Social Studies	.94	.92	.91	.95	.96	1.00

Note: N=801

TABLE 27. Between Content Area Scale Score Correlations at School Level for Grade 8

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.94	1.00				
Lang. Usage	.94	.98	1.00			
Mathematics	.89	.93	.94	1.00		
Science	.95	.94	.94	.95	1.00	
Social Studies	.94	.96	.96	.95	.98	1.00

Note: N=259.

TABLE 28. Number of Items Flagged as Differential Item Functioning for 1998 MSPAP

Grade 3

	Reading (49 items)		Writing (11 items)		Language (32 items)		Math Content (61 items)		Math Process (30 items)		Social Studies (49 items)		Science (53 items)	
	+[1]	-[2]	+	-	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asian	0	1	0	0	0	0	5	0	2	0	1	3	0	0
Hispanic	1	0	0	0	0	0	2	0	1	0	0	0	1	3
Female	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Grade 5

	Reading (54 items)		Writing (11 items)		Language (32 items)		Math Content (63 items)		Math Process (31 items)		Social Studies (51 items)		Science (52 items)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asian	1	0	0	0	0	0	3	0	1	0	1	0	1	0
Hispanic	0	0	0	0	0	0	1	0	0	0	0	1	3	1
Female	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Grade 8

	Reading (59 items)		Writing (12 items)		Language (25 items)		Math Content (48 items)		Math Process (24 items)		Social Studies (57 items)		Science (69 items)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asian	0	0	0	0	0	0	0	0	0	0	0	0	2	2
Hispanic	0	2	1	0	0	0	0	0	0	0	0	1	0	2
Female	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Note 1: The minority group members did better than was expected
Note 2: The minority group members did less well than was expected

TABLE 29. Outcome Difficulty Indicators for each Grade for the 1998 MSPAP

Outcome Number	Outcome	Grade3	Grade5	Grade8
Reading				
2.	Reading for Literary Experience	53	51	59
3.	Reading to be Informed	54	51	46
4.	Reading to Perform a Task	44	47	57
Writing				
1.	Writing to Inform	36	48	53
2.	Writing to Persuade	40	49	48
3.	Writing to Express Personal Ideas	37	47	61
Language Usage				
1.	Language In Usage	36	42	51
Mathematics				
1.	Problem Solving	N/A	48	N/A
2.	Communication	31	44	30
3.	Reasoning	29	45	32
4.	Connections	31	41	34
5.	Concepts/Relationships	46	37	32
6.	Measurement/Geometry	52	44	28
7.	Statistics	55	49	36
8.	Probability	42	48	36
9.	Patterns/Relationships	45	N/A	N/A
9.	Patterns/Algebra	N/A	43	35
Science				
1.	Concepts of Science	47	38	38
2.	Nature of Science	41	40	47
3.	Habits of Mind	48	37	46
5.	Processes of Science	45	38	43
6.	Applications of Science	37	30	36
Social Studies				
1.	Political Systems	42	37	47
2.	People/Nation & World	40	43	53
3.	Geography	41	43	50
4.	Economics	41	37	47
5.	Skills and Processes	42	39	50
6.	Valuing Self and Others	37	44	51
7.	Understand/Attitudes	37	47	44

Note: N/A means the outcome is not measured at that grade.
Note: The numbers are percentages of the maximum possible scores.

Appendix A

Test Maps for 1998 MSPAP

MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM

DATES/TIMES* FOR MAY 1998 ADMINISTRATION

GRADE 3 DRAFT TASK MAP

	Tasks By Day Of Testing														
	MONDAY MAY 11			TUESDAY MAY 12			WEDNESDAY MAY 13			THURSDAY MAY 14			FRIDAY MAY 15		
	#	Subject[▲]	Times	#	Subject[▲]	Times	#	Subject[▲]	Times	#	Subject[▲]	Times	#	Subject[▲]	Times
A	3077	R/M/LWP	100	3078 3063 3079	R M SS	50 30 30	3079 3054 3078	SS M EWP	30 30 40	3078 3080	EWP SCI	55 50	3081	SS/SCI/ LWP SURVEY	85 20
B	3082 3076	M SS/LWP SURVEY	50 45 10	3065	R/LWP/ SCI/SS	110	3065 3083	R/SCI/SS SS	70 35	3036 3065	M EWP/SCI	50 55	3065 3075	EWP SCI SURVEY	55 40 10
C	3084	SCI/M/ LWP SURVEY	95 10	3085	R/SCI	105	3085 3059 3086	EWP M SS	40 35 30	3086 3085 3087	SS EWP M	30 55 25	3074	R/SS/ LWP SURVEY	95 10

▲ Language usage activities are distributed throughout and therefore not listed. Check your Examiner's Manual to determine where they occur.

* Each day is approximately 1 hour + 45 minutes of engaged testing and does not include time for organizing and preparing students for test administration.

BEST COPY AVAILABLE

VERSION DATE: 8-19-97

MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM

DATES/TIMES* FOR MAY 1998 ADMINISTRATION

GRADE 5 DRAFT TASK MAP

	Tasks By Day Of Testing														
	MONDAY MAY 4			TUESDAY MAY 5			WEDNESDAY MAY 6			THURSDAY MAY 7			FRIDAY MAY 8		
	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times
A	5074	SS/R/LWP SURVEY	95 10	5076 5071	SCI/M/W R/LWP	40 65	5073 5056 5076	SCI M SCI/M	55 30 20	5076	SCI/ EWP/SS SURVEY	95 10	5076 5077	EWP M	55 50
B	5079 5081 5058	R M M	55 25 30	5078 5079	R/SCI/ LWP EWP	65 40	5079 5080	EWP SCI	55 50	5078 5072	SCI SS SURVEY	45 45 10	5072 5069	SS M/LWP SURVEY	50 45 10
C	5075 5082	SCI/R SS/LWP	75 35	5075	SCI/R/M	105	5075 5083	EWP SCI/LWP SURVEY	40 55 10	5075 5070	EWP M SURVEY	55 40 10	5084 5018	SS M	70 30

▲ Language usage activities are distributed throughout and therefore not listed. Check your Examiner's Manual to determine where they occur.

• Each day is approximately 1 hour + 45 minutes of engaged testing and does not include time for organizing and preparing students for test administration.

BEST COPY AVAILABLE

VERSION DATE: 8-19-97

MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM

DATES/TIMES* FOR MAY 1998 ADMINISTRATION

GRADE 8 DRAFT TASK MAP

	Tasks By Day Of Testing														
	MONDAY MAY 11			TUESDAY MAY 12			WEDNESDAY MAY 13			THURSDAY MAY 14			FRIDAY MAY 15		
	#	Subject[▲]	Times	#	Subject[▲]	Times	#	Subject[▲]	Times	#	Subject[▲]	Times	#	Subject[▲]	Times
A	8069	SCI/R/ M/SS	105	8069 8070	R/SCI M/SS SS/LWP	60 40	8067 8069	R/LWP EWP SURVEY	60 40 10	8069 8061	EWP SS SURVEY	55 35 10	8029 8071	M SCI	60 45
B	8024 8062	M R/LWP SURVEY	40 60 10	8077	M/SCI/SS	105	8077 8068 8072	EWP SS M/LWP	40 30 35	8077 8073	EWP SS	55 45	8063	R/SCI SURVEY	95 10
C	8064 8060	SS/LWP M SURVEY	65 30 10	8066	R/LWP/ SCI/SS	105	8066 8074	R/LWP/ SCI/SS R	50 55	8075 8074	M EWP SURVEY	55 40 10	8074 8076	EWP SCI	55 50

▲ Language usage activities are distributed throughout and therefore not listed. Check your Examiner's Manual to determine where they occur.

* Each day is approximately 1 hour + 45 minutes of engaged testing and does not include time for organizing and preparing students for test administration.

BEST COPY AVAILABLE

VERSION DATE: 8-19-97

Appendix B

Number of Items Comprising Each Outcome for 1998 MSPAP

Appendix B

Number of Measures for Each Outcome—Grade 3

Cluster A	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	0	6	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	4	5	7	4	6
Math Process	0	10	8	5	0	0	0	0	0
Social Studies	4	4	6	0	7	4	3	0	0
Science	4	5	4	0	4	4	0	0	0

Cluster B	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	6	0	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	9	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	6	5	4	0	4
Math Process	0	4	1	4	0	0	0	0	0
Social Studies	0	5	4	4	6	3	3	0	0
Science	11	6	5	0	5	7	0	0	0

Cluster C	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	0	7	6	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	9	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	9	4	5	4	4
Math Process	0	8	9	4	0	0	0	0	0
Social Studies	5	5	0	5	7	4	4	0	2
Science	6	4	5	0	6	4	0	0	0

Note: See Table 29 for the outcome name corresponding to each outcome number.

Return

BEST COPY AVAILABLE

Appendix B

Number of Measures for Each Outcome—Grade 5

Cluster A	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	6	0	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	0	5	4	6	4
Math Process	0	5	4	3	0	0	0	0	0
Social Studies	4	0	5	6	7	4	6	0	0
Science	5	4	4	0	4	4	0	0	0

Cluster B	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	0	6	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	6	0	4	4	4
Math Process	0	9	9	4	0	0	0	0	0
Social Studies	0	4	6	5	5	5	3	0	0
Science	5	5	4	0	5	4	0	0	0

Cluster C	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	0	6	6	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	9	4	7	4	7
Math Process	0	3	8	4	0	0	0	0	0
Social Studies	5	5	0	7	5	4	3	0	0
Science	7	7	6	0	4	5	0	0	0

Note: See Table 29 for the outcome name corresponding to each outcome number.

Return

BEST COPY AVAILABLE

Appendix B

Number of Measures for Each Outcome—Grade 8

Cluster A

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	0	7	6	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	10	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	9	5	4	0	6
Math Process	0	4	6	5	0	0	0	0	0
Social Studies	5	5	0	4	10	4	4	0	0
Science	8	5	5	0	5	4	0	0	0

Cluster B

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	0	7	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	2	6	4	4	5
Math Process	0	8	5	7	0	0	0	0	0
Social Studies	5	4	9	0	9	4	4	0	0
Science	5	9	4	0	8	4	0	0	0

Cluster C

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	7	5	0	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	5	1	4	4	4
Math Process	0	5	6	1	0	0	0	0	0
Social Studies	4	0	5	5	8	5	4	0	0
Science	5	4	5	0	4	7	0	0	0

Note: See Table 29 for the outcome name corresponding to each outcome number.

Return

BEST COPY AVAILABLE

Appendix C

Scaled Score Ranges for Each Proficiency Level in MSPAP`

Appendix C — Scaled Score Ranges for each Proficiency Level in MSPAP

MSPAP Proficiency level scale score ranges

	Grade		
Level	3	5	8
READING			
1	620 – 700	620 – 700	620 – 700
2	580 – 619	580 – 619	580 – 619
3	530 – 579	530 – 579	530 – 579
4	490 – 529	490 – 529	490 – 529
5	350 – 489	350 – 489	350 – 489
WRITING			
1	614 – 700	- - - -	- - - -
2	577 – 613	567 – 700	551 – 700
3	528 – 576	522 – 566	505 – 550
4	350 – 527	488 – 521	350 – 504
5	- - - -	350 – 487	- - - -
LANGUAGE USAGE			
1	620 – 700	597 – 700	- - - -
2	576 – 619	567 – 596	565 – 700
3	521 – 575	533 – 566	509 – 564
4	350 – 520	350 – 532	474 – 508
5	- - - -	- - - -	350 – 473
MATHEMATICS			
1	626 – 700	617 – 700	618 – 700
2	583 – 625	575 – 616	579 – 617
3	531 – 582	520 – 574	525 – 578
4	489 – 530	473 – 519	481 – 524
5	350 – 488	350 – 472	350 – 480
SCIENCE			
1	619 – 700	625 – 700	619 – 700
2	580 – 618	580 – 624	576 – 618
3	527 – 579	525 – 579	532 – 575
4	488 – 526	484 – 524	482 – 531
5	350 – 487	350 – 483	350 – 481
SOCIAL STUDIES			

BEST COPY AVAILABLE

1	622 – 700	619 – 700	620 – 700
2	580 – 621	580 – 618	582 – 619
3	525 – 579	529 – 579	530 – 581
4	495 – 524	350 – 528	495 – 529
5	350 – 494	- - - -	350 – 494

Dashes indicate proficiency levels for which cut scores could not be established for MSPAP. These cut scores will be established on future editions of MSPAP.

Return

BEST COPY AVAILABLE